

IJCSIS Vol. 9 No. 1, January 2011
ISSN 1947-5500

International Journal of Computer Science & Information Security

© IJCSIS PUBLICATION 2011

Editorial

Message from Managing Editor

The current January 2011 issue of International Journal of Computer Science and Information Security (IJCSIS) continues the tradition of encompassing a wide array of subjects in the fields computer science and information security, ranging from network technologies to security systems. We are very pleased to confirm that IJCSIS is indexed and abstracted in EBSCO and ProQuest, with pending decisions from a number of other database vendors.

Field coverage includes: security infrastructures, network security: Internet security, content protection, cryptography, steganography and formal methods in information security; multimedia systems, software, information systems, intelligent systems, web services, data mining, wireless communication, networking and technologies, innovation technology and management. (See monthly Call for Papers)

We will do our outmost to maintain the established practice of timely publication of regular, as well as special issues of the journal, and further improve the quality of published papers and the journal itself. It will be a great pleasure for us if the readers, reviewers, and authors recognize our intention and efforts, and continue to support us in this mission.

On behalf of the Editorial Board and the IJCSIS members, we would like to express our gratitude to all authors and reviewers for their hard and high-quality work, diligence, and enthusiasm.

Available at <http://sites.google.com/site/ijcsis/>

IJCSIS Vol. 9, No. 1, January 2011 Edition

ISSN 1947-5500 © IJCSIS, USA.

Abstracts Indexed by (among others):



IJCSIS EDITORIAL BOARD

Dr. Gregorio Martinez Perez

Associate Professor - Professor Titular de Universidad, University of Murcia (UMU), Spain

Dr. M. Emre Celebi,

Assistant Professor, Department of Computer Science, Louisiana State University in Shreveport, USA

Dr. Yong Li

School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China

Prof. Hamid Reza Naji

Department of Computer Engineering, Shahid Beheshti University, Tehran, Iran

Dr. Sanjay Jasola

Professor and Dean, School of Information and Communication Technology, Gautam Buddha University

Dr Riktesh Srivastava

Assistant Professor, Information Systems, Skyline University College, University City of Sharjah, Sharjah, PO 1797, UAE

Dr. Siddhivinayak Kulkarni

University of Ballarat, Ballarat, Victoria, Australia

Professor (Dr) Mokhtar Beldjehem

Sainte-Anne University, Halifax, NS, Canada

Dr. Alex Pappachen James, (Research Fellow)

Queensland Micro-nanotechnology center, Griffith University, Australia

Dr. T.C. Manjunath,

ATRIA Institute of Tech, India.

TABLE OF CONTENTS

1. Paper 31121008: ANNaBell Island: A 3D Color Hexagonal SOM for Visual Intrusion Detection (pp. 1-7)

Chet Langin, Michael Wainer, and Shahram Rahimi
Computer Science Department, Southern Illinois University Carbondale, Carbondale, Illinois, USA

2. Paper 31121009: Multimedia Video Conference Identities Study (pp. 8-12)

Mahmoud Baklizi, Nibras Abdullah, Ali Abdulqader Bin Salem, Sima Ahmadpour, Sureswaran Ramadass
National Advanced IPv6 Centre of Excellence, Universiti Sains Malaysia, Pinang, Malaysia

3. Paper 31121012: A Multi-Purpose Scenario-based Simulator for Smart House Environments (pp. 13-18)

Zahra Forootan Jahromi and Amir Rajabzadeh, Department of Computer Engineering, Razi University, Kermanshah, Iran
Ali Reza Manashty, Department of IT and Computer Engineering, Shahrood University of Technology, Shahrood, Iran

4. Paper 31121032: Carrier Offset Estimation for MIMO-OFDM Based on CAZAC Sequences (pp. 19-23)

Dina Samah, Dr. Sherif Kishk, Dr. Fayez Zaki
Department of electronics and communications engineering, Mansoura University, Egypt

5. Paper 31121042: A New Approach to Prevent Black Hole Attack in AODV (pp. 24-29)

M. R. Khalili Shoja, Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
Hasan Taheri, Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
Shahin Vakiliinia, Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

6. Paper 31121047: Application of Web Server Benchmark using Erlang/OTP R11 and Linux (pp. 30-34)

A. Suhendra, A.B. Mutiara,
Faculty of Computer Science and Information Technology, Gunadarma University, Jl. Margonda Raya No.100, Depok 16464, Indonesia

7. Paper 31121021: Advanced Virus Monitoring and Analysis System (pp. 35-38)

Fauzi Adi Rafrastara, Faculty of Information and Communication Technology, University of Technical Malaysia Melaka, Melaka, Malaysia
Faizal M. A, Faculty of Information and Communication Technology, University of Technical Malaysia Melaka, Melaka, Malaysia

8. Paper 23121004: A New Approach for Clustering Categorical Attributes (pp. 39-43)

Parul Agarwal, Department Of Computer Science, Jamia Hamdard (Hamdard University), New Delhi =110062 ,India

M. Afshar Alam, Department Of Computer Science, Jamia Hamdard (Hamdard University), Jamia Hamdard (Hamdard University), New Delhi =110062 ,India

Ranjit Biswas, Manav Rachna International University, Green Fields Colony, Faridabad, Haryana 121001

9. Paper 31121007: Position based Routing Scheme using Concentric Circular Quadrant Routing Protocol in mobile ad hoc network (pp. 44-47)

Upendra Verma, M.E. (Research Scholar), Shri Vaishnav Institute Of Science and Technology Indore (Madhya Pradesh), INDIA

Mr. Vijay Prakash, Asst. Prof., Shri Vaishnav Institute Of Science and Technology Indore (Madhya Pradesh), INDIA

10. Paper 31121010: Comparative Analysis of Techniques for Eliminating Spam in Fax over IP (pp. 48-52)

Manju Jose, Research Scholar, Mother Teresa Women's University, Kodaikanal, India

Dr. S.K. Srivatsa, Senior Professor, St. Joseph College of Engineering, Chennai, India.

11. Paper 31121011: Resource Estimation And Reservation For Handoff Calls In Wireless Mobile Networks (pp. 53-59)

K. Venkatachalam, Professor - Department of ECE, Velalar College of Engineering and Technology, Thindal post, Erode – 638012 , Tamilnadu, India.

P. Balasubramanie, Professor - Department of CSE, Kongu Engineering College, Erode, Tamilnadu, India

12. Paper 31121015: Low Complexity Scheduling Algorithm for Multiuser MIMO System (pp. 60-67)

Shailendra Mishra, Kumaon Engineering College, Dwarahat, Uttrakhand ,India

D.S. Chauhan, Uttrakhand Technical University, Dehradun, Uttrakhand, India

13. Paper 31121020: Email Authorship Identification Using Radial Basis Function (pp. 68-75)

A. Pandian, Asst. Professor (Senior Grade), Department of MCA, SRM University, Chennai, India

Dr. Md. Abdul Karim Sadiq, Ministry of Higher Education, College of Applied Sciences, Sohar, Sultanate of Oman

14. Paper 31121022: Performance of Iterative Concatenated Codes with GMSK over Fading Channels (pp. 76-85)

Labib Francis Gergis, Misr Academy for Engineering and Technology, Mansoura, Egypt

15. Paper 31121025: Priority Based Mobile Transaction Scheme Using Mobile Agents (pp. 86-91)

J.L. Walter Jeyakumar, R.S.Rajesh, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, INDIA.

16. Paper 31121027: Design of Content-Oriented Information Retrieval by Semantic Analysis (pp. 92-97)

S. Amudaria, Dept of IT, SSN College of Engineering, Chennai, India

S. Sasirekha, Dept of IT, SSN College of Engineering, Chennai, India

17. Paper 31121035: Enhanced Load Balanced AODV Routing Protocol (pp. 98-101)

Iftikhar Ahmad and Humaira Jabeen

Department of Computer Science, Faculty of Basic and Applied Sciences, International Islamic University Islamabad, Pakistan

18. Paper 31121037: Comparative Analysis of Speaker Identification using row mean of DFT, DCT, DST and Walsh Transforms (pp. 102-107)

*Dr. H B Kekre, Senior Professor, Computer Department, MPSTME, NMIMS University, Mumbai, India
Vaishali Kulkarni, Associate Professor, Electronics & Telecommunication, MPSTME, NMIMS University, Mumbai, India*

19. Paper 31121041: Performance Evaluation of Space-Time Turbo Code Concatenated With Block Code MC-CDMA Systems (pp. 108-115)

*Lokesh Kumar Bansal, Department of Electronics & Comm. Engg., N.I.E.M., Mathura, India
Aditya Trivedi, Department of Information and Comm. Technology, ABV-IIITM, Gwalior, India*

20. Paper 31121046: An Unsupervised Feature Selection Method Based On Genetic Algorithm (pp. 116-120)

*Nasrin Sheikhi, Amirmasoud Rahmani, Mehran Mohsenzadeh
Department of computer engineering, Islamic Azad University of Iran Research and Science Branch Ahvaz, Iran
Reza Veisisheikhrobat, National Iranian South Oil Company (NISOC), Ahvaz, Iran*

21. Paper 31121050: A PCA Based Feature Extraction Approach for the Qualitative Assessment of Human Spermatozoa (pp. 121-126)

*V. S. Abbiramy, Department of Computer Applications, Velammal Engineering College, Chennai, India
Dr. V. Shanthy, Department of Computer Applications, St. Joseph's Engineering College, Chennai, India*

22. Paper 31121053: Analysis of Error Metrics of Different Levels of Compression on Modified Haar Wavelet Transform (pp. 127-133)

*T. Arumuga Maria Devi, Assistant Professor, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli . TamilNadu.
S. S. Vinsley, Student IEEE Member, Guest Lecturer, Manonmaniam Sundaranar University, Centre for Information Technology and Engg, Manonmaniam Sundaranar University, Tirunelveli. TamilNadu.*

23. Paper 31121056: Information Security and Ethics in Educational Context: Propose a Conceptual Framework to Examine Their Impact (pp. 134-138)

*Hamed Taherdoost, Islamic Azad University, Islamshahr Branch, Department of Computer, Tehran, Iran
Meysam Namayandeh, Islamic Azad University, Islamshahr Branch, Department of Computer, Tehran, Iran
Neda Jalaliyoon, Islamic Azad University, Semnan Branch, Department of Management, Semnan, Iran*

24. Paper 31121038: Finding Fuzzy Locally Frequent Itemsets (pp. 139-146)

*Fokrul Alom Mazarbhuiya, College of Computer Science, King Khalid University, Abha, Saudi Arabia
Md. Ekramul Hamid, College of Computer Science, King Khalid University, Abha, Saudi Arabia*

25. Paper 31121039: An Extensible Cloud Architecture Model for Heterogeneous Sensor Services (pp. 147-155)

*R.S. Ponmagal, Dept of CSE, Anna University of Technology, Tiruchirappalli, India
J. Raja, Dept of ECE, Anna University of Technology, Tiruchirappalli, India*

26. Paper 31121045: Computer Simulation tool for Learning Brownian Motion (pp. 156-158)

*Dr Anwar Pasha Abdul Gafoor Deshmukh
Lecturer, College of Computer and Information Technology, University of Tabuk , Tabuk., Saudi Arabia*

27. Paper 31121052: Mobility Management Techniques to Improve QoS In Mobile Networks Using Intelligent Agent Decision Making Protocol (pp. 159-165)

*Selvan C, Department of Computer Science and Engineering Government College of Technology, Coimbatore, Tamil Nadu, India
Dr. R. Shanmugalakshmi, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamil Nadu, India*

28. Paper 31121060: Security Risks and Modern Cyber Security Technologies for Corporate Networks (pp. 166-170)

*Wajeb Gharibi, College of Computer Science and Information Systems, Jazan University, Jazan, Saudi Arabia.
Abdulrahman Mirza, Center of Excellence in Information Assurance (CoEIA), King Saud University, KSA.*

29. Paper 31121068: Analysis of Data Mining Visualization Techniques Using ICA AND SOM Concepts (pp. 171-180)

*K. S. Rathnamala, Research Scholar of Mother Teresa Women's University, Kodaikanal
Dr. R. S. D. Wahida Banu, Professor & Head, Dept. of Electronics & Communication Engg., GCE.*

30. Paper 31121017: Mobile Agent Computing (pp. 181-187)

Mrigank Rajya, Software Engineer, HCL Technologies Ltd. Gurgaon, India

31. Paper 31121034: Map Reduce for DC4.5 and Ensemble Learning In Distributed Data Mining (pp.188-192)

*Dr. E. Chandra, Research Supervisor and Director, Department Of Computer Science, D J Academy for Managerial Excellence, Coimbatore, Tamilnadu, India.
P. Ajitha, Research Scholar and Assistant Professor, Department Of Computer Science, D J Academy for Managerial Excellence, Coimbatore, Tamilnadu, India*

32. Paper 31121058: A Novel E-Service for E-Government (pp. 193-200)

*A. M. Riad, Hazem M. El-Bakry, and Gamal H. El-Adl
Dept. of Information Systems, Faculty of Computer Science and Information Systems, Mansoura University, Mansoura, Egypt*

ANNaBell Island: A 3D Color Hexagonal SOM for Visual Intrusion Detection

Chet Langin, Michael Wainer, and Shahram Rahimi

Computer Science Department
Southern Illinois University Carbondale
Carbondale, Illinois, USA

Abstract—Self-Organizing Maps (SOM) are considered by many to be *black boxes* because the results are often non-intuitive. Our research takes the multidimensional output from a successful intrusion detecting SOM and displays it in novel full color and 3D formats, with landscape features similar to an island, that assist in understanding the SOM results. This paper describes the visual data mining from the map and explains the methodology in obtaining the full color and 3D maps.

Keywords: Data Mining; Forensics; Intrusion Detection; Modeling; Self-Organizing Map (SOM); Visualization.

I. INTRODUCTION

Self-Organizing Maps (SOM) have been researched for years as being possible methods of intrusion detection. SOM can display multidimensional data in lower dimensions, but the results are often not intuitive, resulting in SOM sometimes being called a *black box* method, meaning that the inner workings are not visible. Security technicians appear to be reluctant to use methods that they do not understand.

SOM methods are actually programmed by design and the creators know exactly what is *inside the box*. The results mystify many technicians, though, resulting in the *black box* epithet. Our visual approach attempts to present the output of a successful SOM intrusion detector in a way that is more comprehensible to people that need to understand how this type of intrusion detection works.

Compare SOM to a hound dog with a good sense of smell. One knows when the dog uses this sense of smell to find something, even if the exact smell is not known. Likewise with SOM, the method can be successfully used, even if the inner workings are not exactly understood by the technicians. The value of our research is that it can help to convince technicians to use SOM as a valid means of intrusion detection instead of dismissing it as being something that cannot be understood, thus helping to find more intrusions.

Previously existing methods of showing SOM in hexagonal formats are discussed in Section II, Related Works. The background of our research leading up to this paper is explained in Section III, Background. How the full color and 3D maps were created is given in Section IV, Methodology, and Section V is the Conclusion.

II. RELATED WORKS

A. Intrusion Detection Development

Amoroso [1] said intrusion detection is the process of identifying and responding to malicious activity targeted at computing and networking sources. Applied intrusion detection was first notably methodized in 1986 by Denning [2]. One of the first published systems was reported by Lunt [3] in the late 1980's and was called the Intrusion Detection Expert System (IDES). It used expert systems and statistics.

The following papers summarized the development of computational intelligent methods in intrusion detection. The use of soft computing methods in intrusion detection was noted by Garcia [4] in 2000. A comprehensive survey of intrusion detection systems was written by Lazarevic [5] in 2005. A comprehensive summary of unsupervised learning algorithms for intrusion detection systems was written by Zanero [6] in 2008. The state of the art of using soft computing methods for intrusion detection was written in 2010 by Langin [7].

B. Using Visual SOM for intrusion detection

The idea for Self-Organizing Maps was developed over a period of years by Kohonen starting in 1976 with his current form conceived in 1982 [8]. (See Kohonen [8] for detailed information about the SOM.) SOM was suggested as a possible method for intrusion detection in 1990 by Fox [9].

Graphical representations of SOM are data mining in the sense that the multidimensional data needs to be organized and displayed in ways that are clearer and can be interpreted.

Fig. 1 shows an early method of visually displaying the information contained in a SOM. It is a 4x4 sample cutout of an 8x8 SOM from Girardin [10] in 1998 which was trained with firewall logs. Each SOM node is represented by a square which is subdivided into four triangular parts with colors and

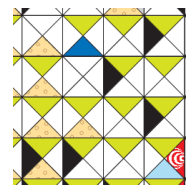


Figure 1,
Pre-Hexagonal

textures to indicate characteristics of that node, resulting in a non-intuitive cryptic display. (A key was not provided for the colors and textures). An alternative layout from the same paper labels each node with an acronym of its primary characteristic, such as http or udp. For example, the upper left node in Fig. 1 was subsequently labeled

with http as being the primary characteristic. See that paper for the entire graphic and an explanation of it.

Fig. 2 shows an 11x3 sample cutout of an 18x14 SOM from Hoglund [11] in a hexagonal format representing user behaviors such as CPU times, characters transmitted, and blocks read in a U-matrix display, meaning that every other hexagon is a node (marked in Fig. 2 with either a dot or a numerical label) and that the intervening hexagons are in a grey scale indicating the distances between the neighboring nodes, darker meaning a larger distance and lighter meaning a closer distance. The labels indicate a user number and the number of Best Matching Node (BMN) hits in a node for that user. For example, 127_8 means that User 127 had 8 BMN hits on the node with that label. A single user as reported in this paper can have hits on nodes in numerous areas of the map. The hexagons provide better representation than a standard 2D layout, but the rectangular layout of the hexagons limits this potential. The U-matrix display is an advantage in that it visually highlights clusters of nodes. The researchers on this project probably have a good idea of the characteristics of various clusters, but these characteristics are not readily apparent from the displayed map. See that paper for the entire graphic and explanation of it. Rectangular U-matrix hexagonal maps were also used by Cho [12] in 2002.

Fig. 3 is a sample cutout of a SOM from Kayacik [13] in 2003 based on network traffic where each hexagon is a node and the amount of filling in the hexagon represents how many BMN hits the corresponding node has (the more hits, the larger the filling). This can create different patterns for different types of traffic, attack vs. normal traffic, for example. See that paper for the entire graphic and explanation of it. A similar histogram map was used by Yeloglu [14] in 2007. This type of map produces useful visual patterns, but does not indicate distances between nodes nor characteristics of nodes.

Fig. 4 is a sample cutout of a U-matrix SOM from Kayacik [15] in 2006 which has been labeled with acronyms and with boundaries drawn to enclose clusters. MHP, for example, stands for *multihop*, and is in a region in this cutout called *host-based attack group*. See that paper for the full graphic and an explanation of it. This type of map provides more information than previous ones, but is still somewhat cryptic.

III. BACKGROUND

This research evolved from a one dimensional SOM, now called 1D ANNaBell, reported by Langin [16 and 17]. This 1D ANNaBell has discovered numerous real life instances of malicious network traffic, being the first self-trained computational intelligence to find feral malware, as far as the authors know, on March 29, 2008, and is still in production after more than two years.



Figure 2, U-matrix

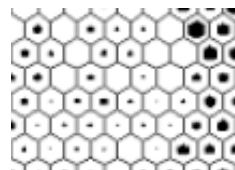


Figure 3,
Histogram

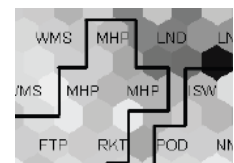


Figure 4,
Acronyms

The 1D ANNaBell was only conceptually a map because there was very little visually, just a jagged line, to observe. A particular jag in the line represented the BMN only for the IP addresses of two local computers infected with a certain kind of bot. This node was used to find additional infected computers. The SOM *hound dog* had the scent, but it was not clear to technicians what the scent was---thus, the *black box* effect.

So 1D ANNaBell was redesigned, using the same data, as a hexagonal map with the intent of producing something visual which would aid technicians in understanding the SOM process. Some of the methodology, using grey scale, for this hexagonal ANNaBell was described by Langin [18 and 19]. This paper continues the methodology by showing how colorization influenced the map, and this paper also shows the map as a 3D *island*. Look ahead to Fig. 14 for the full color map and Fig. 23 for the 3D island to see where this is leading.

The source data for ANNaBell Island is from firewall logs and is in the form of a six dimensional vector for each local IP address---these are the pertinent features, given here as a reference for the rest of this paper:

- 1 tot_norm: *Total normalized*. The total number of log entries in a 24 hour period, normalized. The lowest number of entries in the source data for a local IP address was 0 and the highest number was 2,020,349. These counts were normalized to a range of 0 to 1.
- 2 src_rat: *Source ratio*. The ratio of unique source (external) IP addresses to the total number of log entries.
- 3 port_rat: *Port ratio*. The ratio of unique destination (local) ports to the total number of log entries.
- 4 lo_norm: *Lowest port normalized*. The lowest attempted destination (local) port, normalized from 0 to 1, with the lowest possible port being 0 and the highest possible port being 65,535.
- 5 hi_norm: *Highest port normalized*. The highest attempted destination (local) port, normalized from 0 to 1, with the lowest possible port being 0 and the highest possible port being 65,535.
- 6 udp_rat: *UDP ratio*. The ratio of UDP network traffic to all network traffic.

For example, a local IP address with 1,548 log entries in a 24-hour period, from 139 external IP addresses, directed at 58 local ports, from Port 22 to Port 61,123, with 1,345 of the log entries being for UDP traffic would have a vector of 0.000766204, 0.089793282, 0.0374677, 0.000335698,

0.932677195, 0.868863049.

Over 6,000 (out of 65,536) IP addresses on our local network had no entries in the input data and these were given vectors of 0, 0, 0, 0, 0, 0. This was enough IP addresses to warrant their own node in the SOM, and so a special node was created in the SOM with the vector 0, 0, 0, 0, 0, 0. Since this vector represents the origin in a graph of multidimensional space, this node was called the *Origin*. All other node vectors in the SOM were created with random vectors (as described shortly).

Fig. 5 shows how a meta-hexagonal layout was used instead of a rectangular one because this would allow the SOM nodes to spread out in more directions. It is a large hexagon made of 919 smaller hexagons, with each smaller hexagon representing a node on the SOM. Thus, there is a one-to-one relationship between small hexagons and nodes in this layout. The nodes have numbered labels in a spiral fashion from the center towards the edge. The node with the largest numbered label is Node 918 and is located at the very top. The *island* in ANNaBell Island refers to this meta-hexagon. Other parts of this paper will refer to other features in Fig. 5 later.

Fig. 6 is an enlarged cutout from the center of Fig. 5 and shows how the smaller hexagons, each representing a SOM node, were labeled inside the meta-hexagon. Node 0, the Origin, was placed in the center and the other numbers increased in a clockwise spiral from the Origin. Hexagons 1-6 in yellow indicate nodes with a distance of 1 from Node 0. Likewise, hexagons 7-18 in blue indicate which nodes are a distance of 2 from Node 0 and hexagons 19-36 in green indicate which nodes are a distance of 3. (The colors yellow, blue, and green in this graphic are not related to how these same colors are used in other graphics in this paper.) For comparison, these nodes are a distance of 1 from Node 10: 23, 24, 25, 11, 2, and 9; and, these nodes are a distance of 2 from Node 3: 1, 9, 10, 25, 26, 27, 28, 29, 14, 15, 5, and 6.

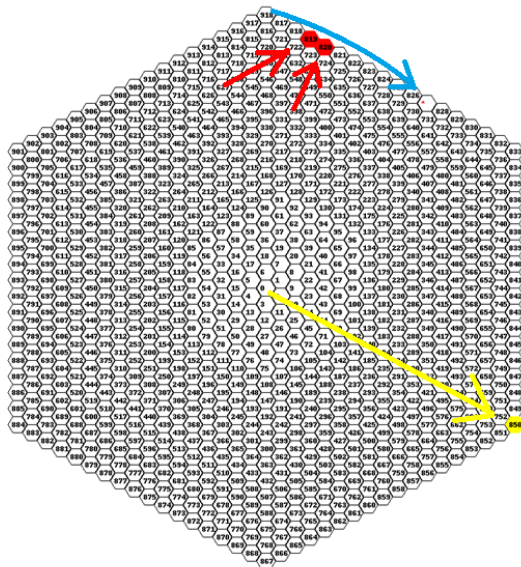


Figure 5, Meta-Hexagonal Layout

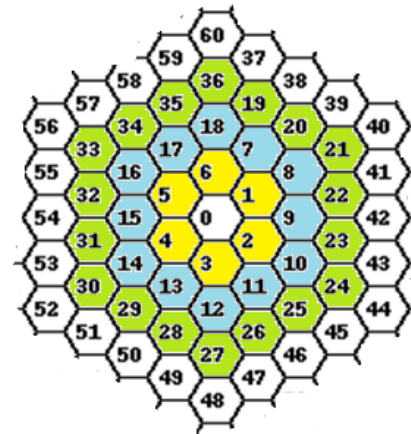


Figure 6, Node Label Numbering Scheme

Some 918 random vectors were created and sorted by their Euclidean distance from the Origin. The closest vector in multidimensional space to Node 0 was assigned to Node 1, the second closest to Node 2, the third closest to Node 3, and so forth up to Node 918, which then had the vector which was furthest away from Node 0 in multidimensional space.

The reason that these vector assignments were made to the nodes in this sorted order was to speed the training time of the SOM by placing at least some of the neighboring nodes in multidimensional space closer to each other in the SOM.

Node movement in multidimensional space was monitored during the SOM training and the training was terminated when the movement stabilized after approximately a week of processing. Referring again to Fig. 5, the Origin moved from Node 0 in the middle to Node 850 in the lower right (yellow). The node furthest from the Origin changed from Node 918 at the top to Node 827 in the upper right (blue arrow and red asterisk). The Best Matching Nodes for the two local bot IP addresses moved from various areas of the map to Nodes 819 and 820 in the upper right (red). (Note that the SOM did not know these were the IP addresses of the bots during the training.) Nodes 819 and 820 were also the BMN for other IP addresses in addition to the IP addresses for the bots. (The colors red, yellow, and blue in Fig. 5 have no relation to how these same colors are used in other graphics in this paper.)

A couple of specific issues and a general issue arose as a result of this training. One specific issue was that 1D ANNaBell did a better job mathematically of isolating the bots (even though ANNaBell Island provided more visual information). Can the hexagonal method be refined to produce as good alerts as the 1D method? The second specific issue was how to represent that nodes (hexagons) 827 and 850 were furthest apart in multidimensional space when they were not furthest apart on the meta-hexagonal island? The general issue was how does one extract other meaningful information from the island? Attempts to answer these questions sparked the methodology reported on below.

IV. METHODOLOGY

Based on the properties of a color wheel, an intuitive hypothesis was made that three of the features could be represented each by the colors of red, green, and blue, and these colors could then be blended for a full color representation of the interaction of these three features.

Fig. 7 shows the basic layout for what was the color experimentation. The *B* in the upper right shows the locations of the BMNs for the bot IP addresses. The *O* in the lower right indicates the location of the Origin. Fig. 8 displays the *tot_norm* values for each hexagonal node scaled in blue. A normalized value of 0 for a hexagon is represented with no blue (white) and a normalized value of 1 is represented by full blue, with other values apportioned in between for various shades of blue. The highest valued hexagon was colored black to distinguish it from the other full blue hexagons (it is 11 nodes (small hexagons) directly above the Origin). The Origin hexagon is appropriately given no blue tint and the nodes (hexagons) representing the bots are relatively dark blue, indicating relatively high *tot_norm* values.

Overall, Fig. 8 shows that the SOM training moved the most active IP addresses toward the upper right edges of the island. Fig. 8 also shows that most IP addresses (most of the island, everything with little or no blue tint), have relatively low *tot_norm* values, i.e., they are not very numerous in the log files.

Fig. 9 shows *udp_rat* shaded in green and Fig. 10 shows *hi_norm* shaded in red. The maximum valued hexagons are colored black so that they can be easily identified.

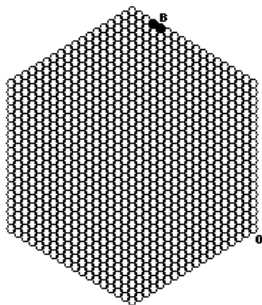


Figure 7, Reference

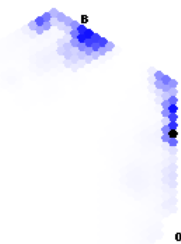


Figure 8, Total Entries



Figure 11, Source Ratio

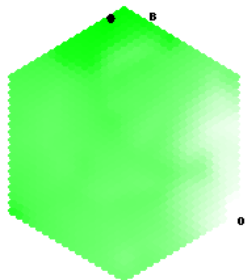


Figure 9, UDP Ratio

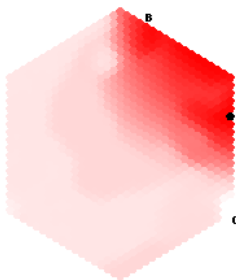


Figure 10, High Ports

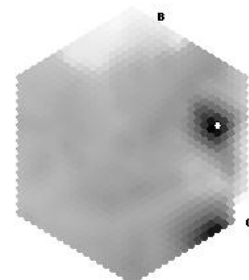


Figure 12, Port Ratio

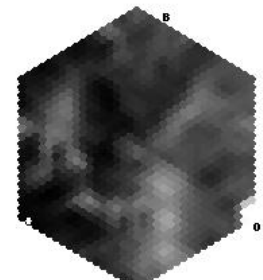


Figure 13, Low Ports

The maximum *udp_rat* hexagon is in the upper left of Fig. 9 and the maximum *hi_norm* is on the right edge in Fig. 10. These figures reveal that most of the IP addresses on the local network have significant UDP traffic and that the SOM moved the high port traffic, generally, from the lowest high ports at the lower left to the highest high ports in the upper right. A pattern has already developed: The most interesting features have been pushed by the SOM to the edges of the island.

Three features still need to be displayed, but there are no more primary colors, so these three additional features were produced in grey scale. (The *tot_norm*, *udp_rat*, and *hi_norm* features were selected for the primary colors because they were suspected of being the most indicative of malicious behavior.)

Fig. 11 is *src_rat* in grey scale. The node with the maximum value is colored white for identification and is in the bottom left of the island. Fig. 12 displays *port_rat* in grey scale with the hexagon containing the maximum value in white. This maximum value is located just inside the edge of the island towards the right and about halfway down. This is the only instance where a maximum value is not on the edge of the island.

Fig. 13 shows the *lo_norm* in grey scale. The node with the maximum value is colored white and is on the lower left edge of the island. This figure is more splotchy than the others which is probably an indication that this feature is not as dominant as the other features.

The next step was to combine the red, green, and blue maps to get a full color representation of those features on the island.

Fig. 14 is a red-green-blue full-color map of the island showing major features. The red was taken from Fig. 10, the green from Fig. 9, and the blue from Fig. 8, all of these three colors being blended together for each hexagon for a full color image. It is now helpful to describe these major features as landscape features to assist in further discussion. The middle right of the map has a dark red tint and is purple with the labels

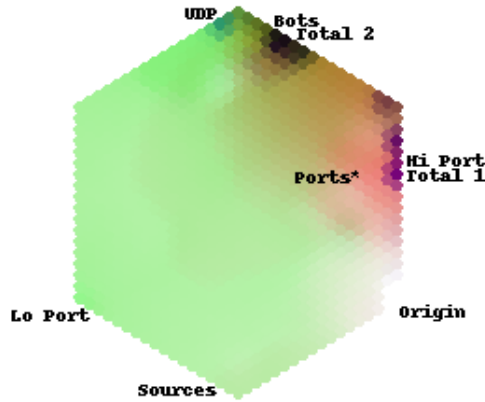


Figure 14, Full Color Map

Ports, Hi Port, and Total 1. Ports refers to the port_rat maximum, Hi Port refers to the hi_norm maximum, and Total 1 refers to the tot_norm maximum. This area of the island was labeled with the landscape designation the Hi Port Mountains (think of purple mountains). The red-tinted area next to it was labeled the Port Cliffs.

The very top of the island in Fig. 14, an area from dark green to black, has the labels UDP, Bots, and Total 2. UDP refers to the area with the maximum UDP ratios and was labeled the UDP Plains. Total 2 refers to an area with secondary high values of tot_norm, so this area of the island was labeled the Bot Hills.

A large part of the island in Fig. 14 is green and was labeled the Valley. It contains areas labeled Lo Port for the lo_norm maximum and Sources for the src_rat maximum. The brown area between the Bot Hills and the Hi Port Mountains was labeled the Plateau. A distance channel image, similar to a U-matrix, was created for comparison with previous methods of analysis.

Fig. 15 shows a distance channel image with the same data and labels as Fig. 14. Unlike a U-matrix, where some hexagons are nodes and other hexagons represent the distances

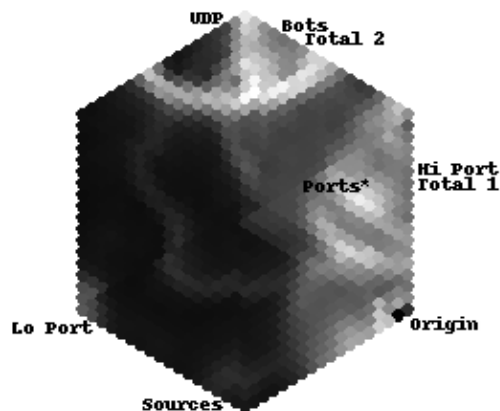


Figure 15, Distance Channel Image

between nodes, in Fig. 15 every hexagon is a node and the shade of grey for that node (hexagon) represents the average distance between a node (hexagon) and the other nodes (hexagons) around it. Dark areas indicate nodes which are closer together and light areas indicate nodes which are farther apart. Imagine farther apart in this context to be similar to elevation. The UDP Plains is clearly delineated as is the Bot Hills. The right side of the island from the Origin to the Port Cliffs and Hi Port Mountains can be seen to be a rocky area with frequent changes in elevations. The Valley is relatively level and the Plateau has a slight elevation. It is not necessary or appropriate to get too technical in evaluating the elevations. This is only an aid in imagining the different parts of the island.

Fig. 16 is a drawing which simplifies the landscape labeling of the island. The Valley is called the Traditional Valley because this is where traditional office network traffic appears, as is further explained below. Origin Basin emphasizes that this is the lowest part of the island.

The next issue addressed was the location of the population on the map. Population in this context means that if the BMNs of all of the local IP addresses were determined, where would they appear on the island? The grey scale method was used, at first, to determine this, and later another method was used. Both will be shown in order.

Fig. 17 plots all (65,536) of the local IP address locations on the island. Dark grey indicates high population for a hexagon (node) and light grey indicates low population. Two areas are relatively highly populated in terms of landscape features: The Valley and the UDP Plains. Fig. 18 displays in light shades of grey the locations of the IP addresses of professionally administered computers, such as desktops for faculty and staff, which are clearly located primarily in the Valley and the Plateau.

A better way of displaying populations was determined. Imagine looking down on Earth from above at night and seeing the lights of villages and cities which indicate populated areas. This was imitated on ANNaBell Island by putting asterisks in hexagons where numerous IP addresses were represented. Red asterisks were arbitrarily used sometimes, and yellow asterisks other times. There is no significant difference between the use of the red and yellow asterisks in showing population locations.

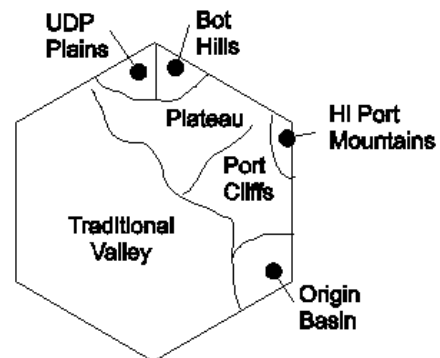


Figure 16, Landscape Drawing

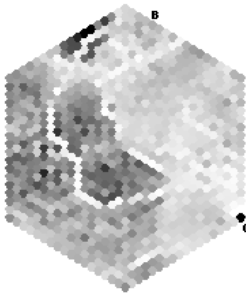


Figure 17,
All IP Addresses



Figure 18,
Well-Kept Computers

Fig. 19 shows the population centers of the IP addresses for the subnet of a well-administered department. The locations are primarily spread out in the Valley with some in the Plateau. Contrast this with locations for IP addresses used by students in Fig. 20, which are largely in the UDP Plains, the Bot Hills, and the Mountains. Showing population centers on the island can clearly be used to characterize the security of various departments, possibly reflecting the skills of the LAN administrators for those areas.

The populated areas of numerous departments were plotted to determine if any differences based upon known security issues could be readily visualized. The results are forensics analyses for organizational departments. Individual IP addresses can also be plotted on the map for an indication of the type of network traffic involved for a single IP address.

Fig. 21 shows the populated area of a department with a history of security problems. Contrast this with Fig. 22, which shows another department which has been locked down by a *paranoid* administrator.

The IP address of any individual computer can be plotted on the map in order to characterize the use of that computer. If an office computer, for example, appears in the UDP Plains instead of the Traditional Valley, then the computer becomes suspect for an infection and/or misuse.

The last step of this research was to display the island in three dimensions. An open source 3D graphics application, Art of Illusion [20], rendered the 3D Island displayed in Fig. 23 by mapping the distance channel data (Fig. 15) to elevation. This

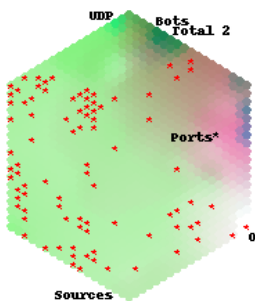


Figure 19,
Well-Kept Department

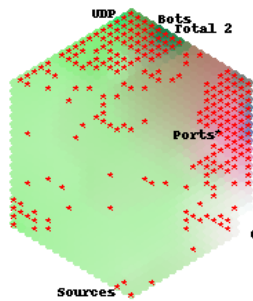


Figure 20,
Student IP Addresses

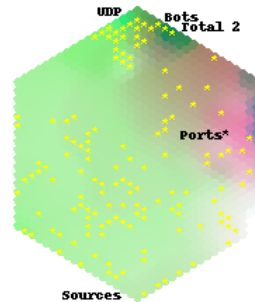


Figure 21,
Poorly-Kept Department

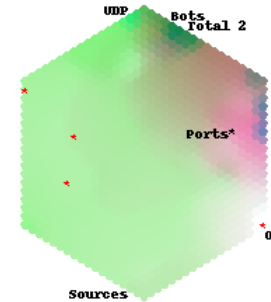


Figure 22,
Paranoid Department

technique, referred to as a height map, is often used in terrain modeling. Here, color was determined based upon height, but colors could have been determined by other data such as that shown in Fig. 14. Indeed, a huge variety of 3D images are possible since any data set or distance channel can be used as an elevation map while colors, textures, and transparencies are supplied by other data sets. To help orient the viewer, Fig. 23 adds background colors of blue for sky and dark blue for water.

V. CONCLUSION AND FUTURE WORK

A hexagonal SOM in a meta-hexagonal layout can graphically display features of network traffic as an island landscape for better understanding of the SOM output, aiding in data mining and forensics and mitigating the *black box* epithet for SOM. This graphical display can also profile networks and individual computers to aid in security and intrusion detection.

This research took the cryptic output of a successful SOM intrusion detector and creatively used color for a 3D landscaped island that represents different types of network traffic, differentiating between malicious and various types of normal behavior. The methodology requires cleverness in manipulating the data channels for visual meaning. Further research in this area would aid in improving informational security intrusion analysis and detection.

There are at least two open questions:

- Would a temporal map maintain the same basic landscape shape or change over time, either randomly or in a meaningful way?
- Is the existing map specific to the tested network or a pattern of Internet traffic, in general?

Much more research can be done in this area, such as the following:

- Track malicious network traffic through the several days leading up to a detection to see if an involved IP address can be seen moving from safe areas to dangerous areas of the map.
- Rebuild the SOM to dynamically handle temporal data, simultaneously training itself and graphically displaying ongoing results.

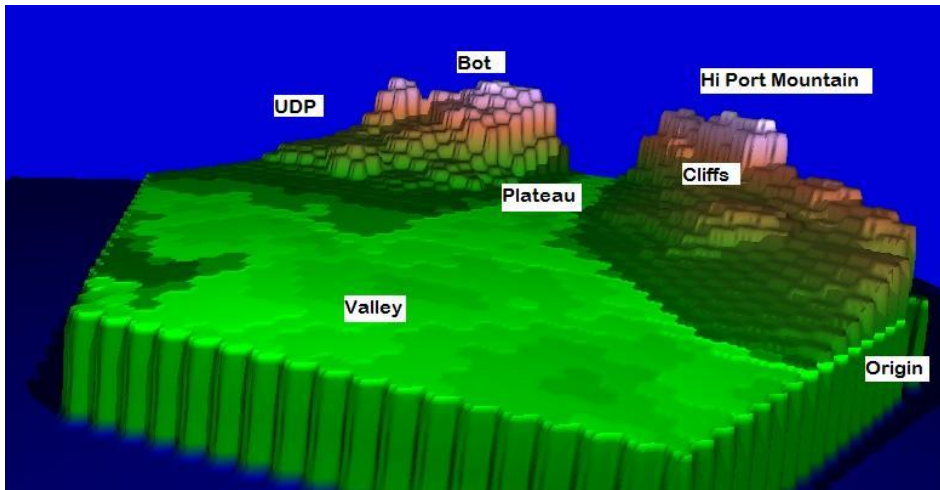


Figure 23, 3D ANNaBell Island

- Create a hierarchical SOM in the Bot Hills area to further differentiate the types of network activity in that area.
- Determine the types of computers that are represented by the Plateau, based on a hypothesis that they are primarily professionally administered servers available to the Internet.
- Address a hypothesis that the UDP Plain represents P2P and/or network gaming traffic.
- An interactive map could be developed giving administrators various tools, such as filters, to aid in visualizing the maps, plus the ability to track changes.

REFERENCES

- [1] Amoroso, E. G., *Intrusion Detection: An Introduction to Internet Surveillance, Correlation, Trace Back, Traps, and Response*, Intrusion.Net Books, 1999.
- [2] Denning, D. E., "An intrusion-detection model." *IEEE Transactions on Software Engineering* 13(2): pp. 118-131, 1986.
- [3] Lunt, T. F., "IDES: An intelligent system for detecting intruders." *Computer Security, Threat and Countermeasures*, 1990.
- [4] Garcia, R. C., and J. A. Copeland, "Soft computing tools to detect and characterize anomalous network behavior," *IEEE Southeastcon*, 2000.
- [5] Lazarevic, A., V. Kumar, and J. Srivastava, "Intrusion detection: a survey." *Managing Cyber Threats*. V. Kumar, J. Srivastava and A. Lazarevic, Springer, pp 19-78, 2005.
- [6] Zanero, S., "Unsupervised learning algorithms for intrusion detection." *Dipartimento di Elettronica e Informazione. Milan, Politecnico di Milano. Dottorato di Ricerca in Ingegneria dell'Informazione*: 163, 2008.
- [7] Langin, C. and S. Rahimi, "Soft computing in intrusion detection: the state of the art." *Journal of Ambient Intelligence and Humanized Computing* 1(2): pp 133-145, 2010.
- [8] Kohonen, T., *Self-Organizing Maps*. Berlin Heidelberg New York, Springer-Verlag, 2001.
- [9] Fox, K. L., R. R. Henning, J. H. Reed, and R. P. Simonian, "A neural network approach towards intrusion detection." *13th National Computer Security Conference*, 1990.
- [10] Girardin, L. and D. Brodbeck, "A visual approach for monitoring logs." *12th Systems Administration Conference (LISA '98)*, Boston, 1998.
- [11] Hoglund, A. J. and K. Hatonen, "Computer network user behavior visualization using Self-Organizing Maps." *International Conference on Artificial Neural Networks (ICANN)*, 1998.

- [12] Cho, S.-B., "Incorporating soft computing techniques into a probabilistic intrusion detection system." *IEEE Trans. Systems Man Cybernet* 32(2): 154, 2002.
- [13] Kayacik, G. H., A. N. Zincir-Heywood, and M. I. Heywood, "On the capability of an SOM based intrusion detection system." *IEEE International Joint Conference on Neural Networks*, 2003.
- [14] Yeloglu, O., A. N. Zincir-Heywood, and M. Heywood, "Growing recurrent Self Organizing Map". *IEEE International Conference on System, Man and Cybernetics, SMC*, 2007.
- [15] Kayacik, H. G. and A. N. Zincir-Heywood, "Using Self-Organizing Maps to build an attack map for forensic analysis." *ACM International Conference on Privacy, Security, and Trust, PST*, 2006.

- [16] Langin, C., H. Zhou, S. Rahimi, "A model to use denied Internet traffic to indirectly discover internal network security problems." *The First IEEE International Workshop on Information and Data Assurance*, Austin, Texas, USA, 2008.
- [17] Langin, C., H. Zhou, B. Gupta, S. Rahimi, and M. Sayeh, "A Self-Organizing Map and its modeling for discovering malignant network traffic." *2009 IEEE Symposium on Computational Intelligence in Cyber Security*, Nashville, TN, USA, 2009.
- [18] Langin, C., D. Che, M. Wainer, and S. Rahimi, "Visualization of network security traffic using hexagonal Self-Organizing Maps." *The 22nd International Conference on Computers and Their Applications in Industry and Engineering (CAINE-2009)*, San Francisco, CA, USA, International Society for Computers and their Applications (ISCA), 2009.
- [19] Langin, C., D. Che, M. Wainer, and S. Rahimi, "SOM with Vulture Fest model discovers feral malware and visually profiles the security of subnets." *International Journal of Computers and Their Applications (IJCA)* 17(4): 1-9, 2010.
- [20] <http://www.artofillusion.org/>, accessed 12/30/2010.

AUTHORS PROFILE

- Chester (Chet) Langin is an Information Security Analyst for Information Technology at Southern Illinois University Carbondale as well as being a Ph.D. candidate there in Computer Science. He has also done soybean bioinformatics research for the university. His research interests include using Soft Computing methods for network intrusion analysis and detection.
- Dr. Michael Wainer obtained his Ph. D. in Computer and Information Science from the University of Alabama at Birmingham, in 1987 and is currently an associate professor of Computer Science at Southern Illinois University Carbondale. His research interests lie in the areas of software development, computer graphics and human computer interaction. He is particularly interested in interdisciplinary work which utilizes the computer as a tool for design and visualization
- Dr. Shahram Rahimi is an associate professor and the director of undergraduate programs at the Department of Computer Science at Southern Illinois University Carbondale. He received his PhD degree from Center for Computational Sciences, Stennis Space Center/University of Southern Mississippi. At the present he is the editor-in-chief for *International Journal of Computational Intelligence Theory and Practice*, the associate editor for *Informatica Journal*, and an editorial board member for several other journals. Dr. Rahimi's research interest includes distributed computing, multi-agent systems, and soft computing. He has over 110 peer reviewed journal articles, proceedings and book chapters in these research areas.

Multimedia Video Conference Identities Study

Mahmoud Baklizi¹, Nibras Abdullah¹, Ali Abdulqader Bin Salem¹, Sima Ahmadpour¹, Sureswaran Ramadass¹.

1: National Advanced IPv6 Centre of Excellence

1: Universiti Sains Malaysia

1: Pinang, Malaysia

1: {mbaklizi, abdullahfaqera, ali, sima, sures}@nav6.org

Abstract— In multimedia conferences there are many Video Conference System (VCS) applications with specific features for each such identity, Support Global Roaming and Support Multiple Client behind NAT. one of the most important features in VCS is the identity, which control and manage the system. The identity feature includes chairman, participants and observers. In this paper we made comparison between the six VCS systems in term of identity feature. At the end we found the NLVC is slightly better than other systems and it is useful for control.

Keywords- NLVC, Eyeball chat, WebEx, Skype, Polycom PVX, Meeting Plaza, Chairman, Participant, Observer.

I. INTRODUCTION

Recently, many areas that used distributed computer network systems are increasing rapidly such as industry, academia, and government. Video conferencing system is a type of computer-based communication applications. The idea of video conferencing appeared for the first time in the 1920s [1]. The task of video conferences focuses to be together in space and time. In addition, it focuses to make groups more effective in their work by using various services such as telephony service over IP networks that are known as IP telephony or Voice over IP [2].

Products of video conference contain a wide range of interactive features of multiple reciprocal actions between the participants. Significantly, in many cases on the internet you can add live video image to a presentation for more effectiveness of education. Supports for a wide range of activities based on video [3].

Voice conversation is the most natural form of interpersonal Communication. In a conversation with at least two participants, every person will finish speaking to be the fact thoughts and hear others transforming. In rare cases, several participants can speak at the same time or one of them can interrupt another, so that occur ambiguous words. Face to Face conversation is, in that all participants in the same physical location, as a meeting place. However, due to the globalization of the activities, there is an increasing need to communicate to a person about geographical locations. Thereby the development of systems becomes to allow a high perception of the

present for face to face like the communication with high-quality speech [4].

The goal of this paper is to investigate how NLVC (Northern Light Videoconference) reacts to network congestion that is, how NLVC manages and handles the sessions to improve slightly comparing with five products Eyeball Chat, Meeting Plaza, Microsoft Communicator, Polycom PVX, and Skype systems.

The paper is organized as follows: in Section II, we present a brief of user identities; in Section III we describe the NLVC system; Section IV describes the comparison of products; Section V shows the comparison of NLVC and other products; finally Section VI concludes the paper.

II. USER IDENTITIES

In general, any conference and meeting session is made up of at least one of the following user identities: (i) conference chairman; (ii) participants; (iii) observers. The conference participants must be existing in any conference meeting.

-The Observer site is allowed only to see the conference, and it is not admitted by active participation in the conference.

-The Participant site can take a contribution and part of the conference. In other words, every participant can send inquiry to an active side to send audio, video, chat and files [5].

-The Chairman site initiates the conference which considers the responsible on the pre-coordination, organization, and company. Moreover, the chairman site can decide, 'why / what', 'if', 'where', and 'that' the meeting. It can also organize and publish an agenda meeting.

The chairman site has the following features:

- To have the ability to cut a short part of a lengthy active site, that is, to 'Kill' the currently site that is activated.

- To have the ability to change the status of a participant to observer and vice-versa.

- To have the ability to terminate or end the conference.

It can become only one side of conferencing leader [6][5].

Generally, a conference is made up of the conference chairman, participants and observers. The conference chairman is the organizer and coordinator of the conference while other conference members can be participants or observers [7][8].

III. NLVC SYSTEM

Desktop Conferencing System has become immensely popular in substituting real meetings and conferences because time cannot be compromised in today's world. The idea of NLVC was adopted by [9]. NLVC depends on control criteria for the optimization of the document and bandwidth of the multimedia Conferencing. The current NLVC implemented by the Network Research Group (NRG) from the school of Computer science at the University Science Malaysia in collaboration with Multimedia Research.

The current NLVC used to utilize a switching method to gain low bandwidth exhaustion. Nowadays, NLVC still allows an unlimited number of users to participate in the conference. NLVC is a set of conference control options that can be considered as rules for controlling the current NLVC that is called Real-time Switching (RSW) control criteria [10]. RSW is used to handle two issues in multimedia conferencing. The first one is to handle the confusion that is generated when everyone tries to speak at the same time. The second issue is the tremendous amount of network traffic that is generated by all participating sites [2]. In addition NLVC has a useful interface as shown in Figure 1.



Fig 1: Screenshot of NLVC program

IV. THE PRODUCT COMPARISON

1- EYEBALL CHAT SYSTEM

The Eyeball chat is a free software parcel, integrated live video, the conversation, the exchange of video-messages and the transference of graphic arts files, and audio. The Eyeball chat offers high-quality video chartrooms with the management of the private sphere that integrates with AIM, messenger MSN and Yahoo instruments compatible [3].

According to the fact that business communication and electronic commerce become more popular and move into cyberspace, the businessmen like to use Eyeball as a result of adopting with technologies changing. [11].

The eyeball is an instrument super massive-shared media where the communication distance allows a different distribution channel, like video, audio, and text messages. The video becomes high-quality, also even in a slow internet connection. Another interesting Feature of this software is the user friendliness.

Eyeball chat can be used either like video-messenger or video communication tool. This last feature gives the user the more permit to send and receive video messages. In addition, it is very useful, especially when the other users are temporarily offline. Moreover, the web version of the software allows the user to begin a chat session directly from the web browser. This feature is useful- for example- the user can access the service from other locations by using the options that offer in Eyeball chat main menu without the chairman handles [12] as shown in Figure 2.

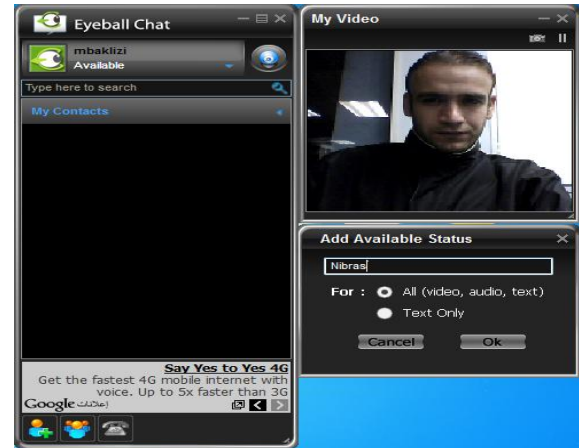


Fig 2: Screenshot of Eyeball Chat main menu

2- WEBEX SYSTEM

WebEx is considered an easy possibility to exchange the ideas with each other everywhere.

WebEx has estimated, that the presidium to conferencing system tools like instant messaging that include chat, audio-conferencing, presence awareness, video conferences, application Sharing, whiteboard, vote and recording of meeting information [13].

WebEx combines the real-time desktop sharing with phone conferencing therefore everybody sees the same thing during speaking by using the sharing controls [14] as shown in Figure 3.



Fig 3: Screenshot of WebEx Chat main menu

It can easily expect the coming of application sharing products like Microsoft Net meeting, Lotus seminal time and WebEx.com come. Participants in the application sharing can allow to the others to see and even control their desktop applications in the same time the chairman can control of their desktop application [15].

WebEx is the famous on-line system meeting for Global business. An efficient communication-tool supports a wide palette of on-line meeting services like application sharing, whiteboard and video conferences. The idea is a build up a number of switching centers that are increasing in the world. Switching centers are responsible for the routing of the communication between end users. This strategy is actually concurrent for a big number of video conferences. Even where two members are in the same area, the data always transmitted from switching centers that will originate an unreasonable delay for video communication [14].

3- SKYPE

Nowadays, the efficient transport of Multimedia-Flows is an open edition and a hot subject as long as multimedia services are rapid in meaning. The Voice-over-IP- applications are taking an ever growing importance like the effectiveness of Skype

application for end users and the application of big networks under SIP[16].

There are several VoIP software applications on the internet as Skype, Google talcum, Windows Live messenger, and Yahoo messenger. After the best knowledge, we insure, only Skype supports multi-party Conferencing [4].

Skype is an interactive software application which allows users to make calls over the internet to different parts of the world. It became popularly under addition of instant messaging, audio/video-entertainment, and file transfer Features. In Skype, phone calls to normal landline and mobile phones need to an emblematic fee based on debiting from the user account system while phone calls to other Skype users in the service are free.

Skype is principal in the educational setting to add interaction to online courses. Students can do chatting with each other or with their teacher using text, audio or video [17].

Skype [18] is a well-known conferencing service with several Million registered users communicate on (best effort) the internet [19]. Figure 4 is shown Screenshot of Skype program.

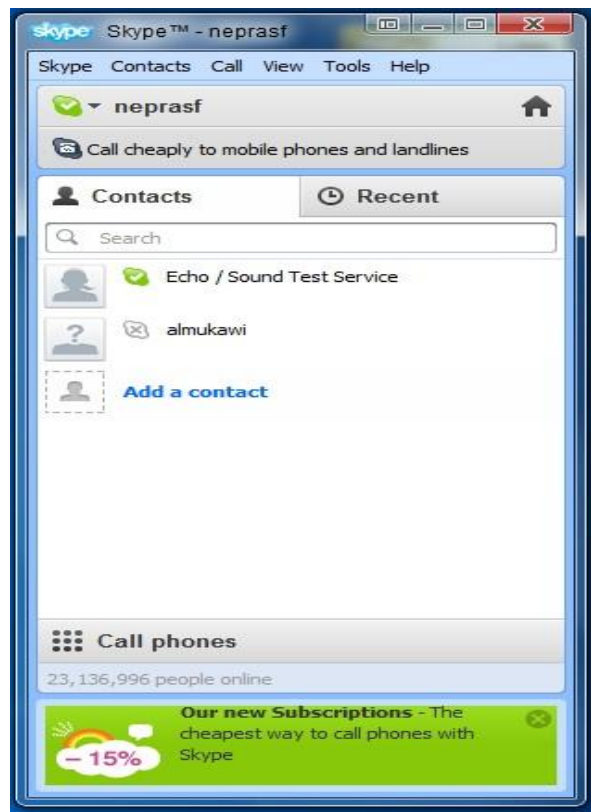


Fig 4: Screenshot of Skype Program

4- POLYCOM PVX

The Polycom PVX video conferences are class economic applications package that delivers qualitatively high-quality audio, video and content sharing on the PC and USB camera. The Polycom PVX application is an excellent visual communications solution for small teams, who do not have dedicated IT support or a need to centrally manage user access or capabilities. PVX System offers high-quality video-phone calls where the pupils are able to see the lecturer and the other pupils in the class. Indeed, is the desktop picture that will transfer is a converted scan with low quality. PVX should be completed with the shared view. In addition, the pupil does not have a text chat characteristic in the PVX system [17]. The Figure 5 is shown the Polycom interface.



Fig 5: Screenshot of Polycom program

5- MEETINGPLAZA

MeetingPlaza is a pure software solution. The most characteristic in MeetingPlaza is "high availability" that no matter how bad the situation can connect and

use. In 17 open trial conducted in the internet has been successfully proven. MeetingPlaza has been adopted as the B/S structure (browser / server) without having to install client software. MeetingPlaza takes up little space and it is able to achieve high-quality and reliability of audio and video communications, document sharing, collaborative browsing, text chatting and other meeting functions, effectively saving time and money, and improving the work efficiency even without the high cost of inputs. MeetingPlaza video conferencing system can be used internal communication, meetings, the exchange of customers and partners in the business, customer service systems, remote training, education system, advisory system, medical tele-consultation system, and many other industries through the network remote real-time audio and video communication systems[20]. Meeting Plaza WCS (Web Contact Service) was improved and connected remote points using VoIP with the internet and PCs for removing the restriction of transporting time and distance. For this reason, the real-time are supported and made in practical to window service for customers [21]. Figure 6 shows the screenshot of Meeting Plaza software.



Fig 6: screenshot of Meeting Plaza software

V. COMPARISON NLVC WITH OTHER PRODUCTS

The five applications are described above and compared to the following three features – chairman, participant, and observer in Table 1.

Table 1: Comparison of multimedia systems

Identity	NLVC	Eyeball Chat	WebEx	Skype	Polycom	Meeting Plaza
Chairman	YES	NO	YES	YES	NO	NO
Participant	YES	YES	YES	YES	YES	YES
Observer	YES	NO	NO	NO	NO	NO

VI. CONCLUSION

This paper compared the performance of NLVC and five technologies in term of user identities –chairman, participants, and observer as the common features for a useful technology. These technologies are Eyeball chat, WebEx, Skype, Polycom PVX and Meeting Plaza. From the discussion and comparison, we conclude that NLVC performs slightly better than other five technologies.

REFERENCES

- [1] E. M. Schooler, "Conferencing and collaborative computing," *Multimedia Systems*. Vol. 4, pp. 210-225, 1996
- [2] B. Mahmoud, F. Nibras, O. Abouabdalla, and A. Sima, "SIP and RSW: A Comparative Evaluation Study," (IJCSIS).Vol. 8,2010
- [3] P. Craven, B. Keppy, and J. Baggaley, "Online video-conferencing products", *The International Review of Research in Open and Distance Learning*, Vol. 3, No. 2, 2002.
- [4] B. Sat, Z. X. Huang, and B. W. Wah, "The design of a multi-party VoIP conferencing system over the Internet," in *Proc. IEEE Int'l Symposium on Multimedia*, Taichung, Taiwan, Dec. 2007, pp. 3-10.
- [5] R. Sureswaran, K. Tharmaraj, H. J. Guyennet, J.C. Lapayre, and H.Y Chan, "A fully distributed architecture to support multimedia conferencing", *Internet Workshop*, 1999. IWS 99, vol., no., pp.261-267,1999
URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=811023&isnumber=17595>
- [6] W. Wang, J. M. Haake and J. Rubart, "Supporting virtual meetings in the overall business context", *International Journal of Computer Applications in Technology* 19(3/4) (2004) 195-208.
- [7] R. Sureswaran, A. Osman, M.S. Mushardin, M. Yusof, and B. Husain, "Advanced Mobile Multipoint Real-Time Military Conferencing System (AMMCS)", In *Proceeding of SSGRR 2002 Conference*. L'Aquila, Italy. July 2002.
- [8] R. Sureswaran, A. Osman, M.S. Mushardin, and M. Yusof, "Advanced Mobile Military Conferencing System (AMMCS)", In *Proceeding of Asia Pacific Defence Electronics (APDEC) 2003*. Kuala Lumpur, Malaysia. 2003.
- [9] R. Sureswaran. "A Control Criteria for Document and Multimedia Conferencing". *Proceedings of IASTED conference on Multimedia and Distributed Systemis*, Honalulu, Aug. 1994
- [10] M.S. KOLHAR, A.F. BAY AN, T.C. WAN, O. ABOUABDALLA, and S. RAMADASS, "Control and Media Sessions: IAX with RSW Control Criteria", *International Conference on Network Applications, Protocols and Services 2008 Executive Development Center*, Universiti Utara Malaysia. pp 75-79.
- [11] A. Winger, "Face-to-Face Communication: Is It Really Necessary in a Digitizing World" in *Business Horizons*, Vol. 48, 247-253, 2005
- [12] R. Giuseppe, and G. Carlo, "Towards CyberPsychology: Mind, Cognitions and Society in the Internet Age", Amsterdam, IOS Press, 2001, 2002, 2003
- [13] C. Bourras, E. Giannaka, and E. Tsiatos, "eCollaboration Concepts, Systems and applications". *Information Science Reference, E-Collaboration: Concepts, Methodologies, Tools, and applications-Vol1, Section I, Chap. 1.2*, N. Kock, Ed.
- [14] <http://www.webex.com/>
- [15] J. Cadiz, A. Balachandran, E. Sanocki, A. Gupta, J. Grudin and G. Jancke "Distance learning through distributed collaborative video viewing", *Computer Supported Cooperative Work (CSCW)*, pp. 2000.
- [16] L. De Cicco, S. Mascolo, and V. Palmisano, "An Experimental Investigation of the Congestion Control Used by Skype VoIP". *WWIC 07*, May 2007.
- [17] F. Martin, and D. Noonan, "Synchronous technologies for online teaching," *Technology for Education (T4E)*, 2010 International Conference on, vol., no., pp.1-4, 1-3 July 2010

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5550062&isnumber=5550032>

[18] Skype, March 2008. <http://www.skype.com>.

[19] A. Petlund, K. Evensen, C. Griwodz, and P. Halvorsen. "TCP mechanisms for improving the user experience for time-dependent thin-stream applications", In *The 33rd Annual IEEE Conference on Local Computer Networks (LCN)*, 2008.

[20] <http://www.meetingplaza.com.cn/intro.php>

[21] S. HIDEYA, and K. SHUJI, "Call Center/Contact Center ASP for Operators Meeting Plaza WCS", *NTT Gijutsu Janaru*, ISSN:0915-2318, VOL.15; NO.8; PAGE.54-58(2003)

<http://sciencelinks.jp/j-east/article/200318/000020031803A0538408.php>



Mahmoud Khalid Baklizi is a researcher pursuing his PhD in Computer Science at the National Advanced IPv6 Center of Excellence in University Sains Malaysia. He received his first degree in Computer Science from Yarmouk University, Jordan, 2002 and his Master degree in Computer Information System from the Arab Academy for Banking and Financial Sciences, Jordan in 2008. His research area of interest includes Multimedia Networking.



Nibras Abdullah Faqera received his Bachelor of Engineering from College of Engineering and Petroleum, Hadhramout University of science and technology, Yemen, 2003. He obtained his Master of Computer Science from School of Computer Science, Universiti Sains Malaysia in 2010. He is academic staff member in Hodeidah University, Yemen. He is researcher pursuing his PhD in Computer Science at the National Advanced IPv6 Center of Excellence in University Sains Malaysia. His research area of interest includes Multimedia Conferencing System (MCS).



Ali Abdulqader Bin Salem: received B.S (computer science) degree from Al-Ahgaif University, Yemen in 2006 and M.S (computer science) from University Science Malaysia (USM), Malaysia in 2009. Currently, he is a PhD student at National Advance IPv6 Center (NAv6), (USM). His current research interests include wireless LAN, multimedia QoS, and video transmission over wireless, distributed system, P2P, and client-server architecture.



Professor Dr. Sureswaran Ramadass:

is a Professor and the Director of the National Advanced IPv6 Centre (NAV6) at Universiti Sains Malaysia. Dr. Sureswaran obtained his BsEE/CE (Magna Cum Laude) and Masters in Electrical and Computer Engineering from the University of Miami in 1987 and 1990 respectively. He obtained his PhD from Universiti Sains Malaysia (USM) in 2000 while serving as a full time faculty in the School of Computer Sciences.

Dr. Sureswaran's recent achievements include being awarded the Anugerah Tokoh Negara (National Academic Leader) for Innovation and Commercialization in 2008 by the Minister of Science and Technology. He was also awarded the Malaysian Innovation Award by the Prime Minister in 2007. Dr. Sureswaran is also the founder and headed the team that successfully took Mlabs Systems Berhad, a high technology video conferencing company to a successful listing on the Malaysian Stock Exchange in 2005. Mlabs is the first, and so far, only university based company to be listed in Malaysia.

A Multi-Purpose Scenario-based Simulator for Smart House Environments

Zahra Forootan Jahromi and Amir Rajabzadeh*

Department of Computer Engineering
Razi University
Kermanshah, Iran

zahra.forootan@gmail.com, rajabzadeh@razi.ac.ir

Ali Reza Manashty

Department of IT and Computer Engineering
Shahrood University of Technology
Shahrood, Iran

a.r.manashty@gmail.com

Abstract: Developing smart house systems has been a great challenge for researchers and engineers in this area because of the high cost of implementation and evaluation process of these systems, while being very time consuming. Testing a designed smart house before actually building it is considered as an obstacle towards an efficient smart house project. This is because of the variety of sensors, home appliances and devices available for a real smart environment. In this paper, we present the design and implementation of a multi-purpose smart house simulation system for designing and simulating all aspects of a smart house environment. This simulator provides the ability to design the house plan and different virtual sensors and appliances in a two dimensional model of the virtual house environment. This simulator can connect to any external smart house remote controlling system, providing evaluation capabilities to their system much easier than before. It also supports detailed adding of new emerging sensors and devices to help maintain its compatibility with future simulation needs. Scenarios can also be defined for testing various possible combinations of device states; so different criteria and variables can be simply evaluated without the need of experimenting on a real environment.

Keywords- smart house simulator; scenario-based smart house; virtual smart house; sensor simulator.

I. INTRODUCTION

As new technologies are emerging, people are more eager to apply these technologies to their house in order to be more and more comfortable and secure. Smart houses, as a state-of-the-art technology in two last decades, are becoming the most exciting and useful tools in our daily lives, which has brought a higher comfort and security level into our life.

The terms smart homes and intelligent homes have been used for more than a decade to introduce the concept of smart devices and equipment in the house. According to the Smart Homes Association the best definition of the smart home technologies is "The integration of technology and services through home networking for a better quality of living".

Smart home is not only an interesting topic, but also a burgeoning industry as well as entering to a broad audience home gradually [1]. Most programmers have to design smart

home systems case by case and spend a lot of time managing them [2]. Many others have already presented how to cut down the building costs by using smart home simulators or high level programming languages [3].

Smart houses could be divided into two main categories:

- Programmable houses – are those scenario-based systems programmed to perform an action triggered by a condition on a sensor output.
- Intelligent houses – are those that possess some kind of intelligence without the need of precise manual design of the procedures.

A. Programmable Houses

Programmable houses will be those that have reactions based only on simple sensor inputs, and possess no built-in intelligence. Such a house for a predefined input has a programmed set of actions to perform.

Examples of such actions might be light bulbs operated by movement sensors, or selection of one of the predefined lighting settings by a button on a remote controller.

Actually, many of currently manufactured and sold smart house systems belong to this group.

The biggest problem with this type of houses is that they have to be reprogrammed when some of the features change. That presents a problem for many people and requires calling a technician to get the job done.

Hence increasing tension to develop some smart home solution that is based on artificial intelligence will adapt its operation to changing user behavior. That tension leads to development of the houses that belong to the second category. This paper, though, supports the first group of smart houses described earlier.

B. Intelligent Houses

They represent the state-of-the-art technology. Those types of installations are driven by artificial intelligence, and instead of having to be programmed they are able to learn basing on observation of inhabitants behavior over a period of time.

*Corresponding Author

One of the first successful implementations was well known Adaptive House developed by M. Mozer at University of Colorado back in 1998. Some other examples that belong to the group of intelligent houses are:

- Georgia Tech Aware Home
- AIRE spaces at MIT
- Interactive Workspaces Project at Stanford
- Gaia project at UIUC
- MavHome project at UTA

The smart house consists of a large and wide ranging set of many services, applications, equipment, networks and systems that act together in delivering the “intelligent” or “connected” home in order to maintain security and control, communications, leisure and comfort, environmental integration and accessibility. These components are represented by many actors that interact and work together to provide interactive systems that benefit the home based user in the smart house. Because of this wide ranging variability of the entities in the smart house, there is a very high level of potential complexity in finding the optimal solution for each different smart house.

For researchers and engineers, it is difficult to work in the real smart home since home appliances are very expensive.

In this paper we present the designing and implementation of a comprehensive smart house simulator to reduce these complexities of implementation a smart house and also find the best solution of making a home or a building smart. Our simulator is completely object based, because we have considered no limitation in different process of simulation.

II. RELATED WORKS

There have been lots of works on this research area including the big corporations and research groups. As a result, various ubiquitous computing simulators such as the Ubiquitous Wireless Infrastructure Simulation Environment (Ubiwise) and TATUS and Context Aware Simulation Toolkit (CAST) have been proposed. The Ubiwise Simulator is used to test computation and communication devices. It has three dimensional (3D) models that form a physical environment viewed by users on a desktop computer through two windows [4, 5]. This simulator focuses on device testing, e.g., in aggregating device functions and exploring the integration of handheld devices and Internet service. Thus, this simulator does not consider an adaptive environment. TATUS is built using the Half Life game engine. Therefore, it looks like an assembled simulation game. It constructs a 3D virtual environment, e.g., a meeting scenario. Using this simulator, a user commands a virtual character to perform tasks, such as to sit down. This simulator does not consider device simulation [6]. CAST is a simulator for the test home domain. This simulator uses scenario based approach. It has been proposed as a prototype using Macromedia's Flash MX 2004 [7]. However, using Flash MX [8] does not support users to freely control their environment. Joon Seok Park et al. proposed the design structure for smart home simulator

regardless of environment factor as well as interaction aspect [9].

III. PROPOSED SMART HOUSE SIMULATOR

There are many simulators in different scope of science and the main purpose of implementing and developing them is demonstrating a virtual model of real subject as well, in order to decrease the problems and difficulties emerge in the way of implementing and evaluating the proposed project in reality.

Indeed researchers use simulators to decrease costs and consumed time for testing and evaluating their ideas on developing and evaluating a project. So the principle duty of a simulator is simulating a virtual model of reality that must be close to its actual model in the real world

In this paper, we present the designing and implementation of smart house simulator for developing and evaluating smart house projects to decrease the obstacles in the way of such projects, mostly cost and time. Due to some difficulties such as providing the necessary real sensors and home appliances to analyze the real home environment, couldn't advance any further than their design level.

This simulator can be used as a substitution for the corresponding real smart environment. Every kind of state-of-the-art sensors and home appliances can be used in the proposed simulator. All the necessary requirements for making a house smart are provided in the simulator.

In the following sections we explain the designing and implementation level of the project and then discuss about the main features of the proposed simulator.

All the principle futures and main capabilities are considered in the designing level, which distinct the proposed simulator from other similar systems.

Some of the most important characteristics of the simulator are describing in the following sections. These principle features of the proposed system are illustrated in Fig.1.

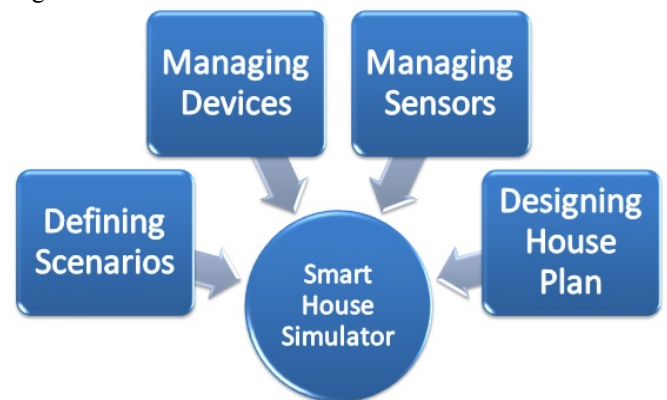


Figure 1: Principle features of the proposed Smart House Simulator

A. Top view plan of the specified house

The simulator should have the capability of demonstrating the plane of the desired house plan in order to be able to simulate a more real virtual model of the house (Fig. 2). The possibility of drawing the house plane is provided in this simulator, so the user can define all



Figure 2: Simulator environment containing top view plan of the house and virtual simulated devices

boundaries of the house such as different rooms, doors, windows and etc. The user also can load an image of top view plan of a house as the house plan.

After designing the house plan user should place each home appliance in their positions as they placed is in real house so user can distinguish them easily for crating different tasks in different objects. User can design the most real model of the real house by using this capability of the proposed simulator.

B. Supporting all kinds of sensors and home appliances

Using different types of sensors and actuators for getting and setting status of each device is an inseparable part of every smart house projects. Many of these sensors are too expensive and some of them have various kinds with different futures of a certain type.

As technology is improving so fast, it's obvious that every day a new kind of sensors, actuators and home appliances will emerge, so the ability of supporting any kind of sensors and actuators is an important future for a smart house simulator.

This simulator provides the possibility of creating a virtual model of any kind of cutting-edge sensors and devices with defining all of their details like the kind of data that each sensor can sense (Fig. 3).

As it is shown in this figure, first user should enter a name for the considered sensor and then select the data

format of it. Data format is the format of the considered sensor that each sensor uses it for demonstrating the status of environment. We have considered 3 data format for sensors contains "Numeral", "Point" and "Multi States", because almost all the sensors data format can be in one of these kinds of data. For example light sensor data format is multi state and it means this sensor use for example two state of "On" and "Off" for showing the light status of environments.

The data format of each sensor can be defined via this form, so that every kind of sensors will all details can be simulated and have a very close model of each sensor in order to have an optimal simulation of smart houses.

For example a light sensor demonstrate the level of light by describing it in 3 level of light, dim and dark; but a temperature sensor show the temperature of an environment in range of numbers or a temperature sensor demonstrates the temperature status of the environments in a range of number, so user should choose the numeral data format for this kind of sensor.

So there is no limitation for using any kind of off the shelf equipment for simulating a virtual smart house environment as well and user can add any number and kind of sensors and home appliances in the simulator. Fig. 4 illustrates the Windows Form which handles adding any virtual model of devices to the simulator. Then user should assign related sensors to each device and choose the devices icon to be shown in designed house environment.

The form is titled 'Add Sensor'. It contains the following fields and controls:

- Sensor Name:** A text input field with the value 'light'.
- Sensor Data Format:** A dropdown menu with 'MultiState' selected.
- Multi State Sensor:** A section containing:
 - Insert State's Name:** A text input field.
 - Sensor Multi States:** A list box containing 'Light', 'Dim', and 'Dark'.
 - Add State** and **Remove State** buttons.
- Added Sensors:** A list box containing 'Power', 'Temperature', 'Location', and 'Light'.
- Add Sensor** button at the bottom right.

Figure 3: Add sensor form. This form enables the user to create any kind of sensors for use in the devices.

The form is titled 'Adding house devices'. It contains the following fields and controls:

- Name of the Object:** A text input field with the value 'TV'.
- Object Icon:** A vertical list of icons representing different house objects.
- List of Added objects:** A list box containing 'Lamp', 'Heater', 'Camera', and 'TV'.
- Add Object's Sensors:** A section containing:
 - Sensors:** A dropdown menu with 'Power(MultiState)' selected.
 - Add Sensor** button.
 - Object's Sensors:** A list box containing 'Power(MultiState)'.
 - Remove Sensor** button.
- Add House Object** button at the bottom.

Figure 4: Adding house devices form

C. Connecting to house remote controlling systems

There are a number of houses remote controlling projects which controlling the devices of a smart house remotely via web or mobile messaging systems.

The proposed simulator can connect to the server of these systems through a network or web. Users can observe the designing simulated house by using internet or via a network and define a task to be done and then send it as a command to the house remote controlling server. The simulator is always checking the server and applies the commands on proper devices.

After each tasks done, the simulator send the updated status of each device to the server. The format of updated status for sending to the server should be in a certain format as Fig. 5 illustrates. This feature enables the use of this simulator as a good substitution for smart houses testing facilities.

Object_ID	Sensor_ID	Sensor_Value	TimeStamp
-----------	-----------	--------------	-----------

Figure 5: The fields in the packet used to send data to smart house remote controlling server

D. Planning scenarios

Scenarios make it easy for people saving the list of actions for further use, in addition to design multiple actions to be done in a single scenario. Later the scenarios can be enabled / disabled in the scenarios list or be used in another scenario too. Cheng, Wang and Chen proposed a reasoning system for smart houses that is also scenario based [10].

One of the capabilities of this simulator, which distinct it from other smart house simulators is the ability of creating scenarios. Each scenario consists of some scheduled tasks and each scheduled task defines a particular action for executing on a special device. As it is shown in figure 5, each scenario has a name and first executing time and date. It means for the first time that a scenario will be executed, user should define the date and time of executing it. A repeating time is considered to repeat the scenario automatically after the first time it executed, and there is no need that user each time Enable the scenario.

So by using a scenario, a set of tasks will execute continuously. As it is shown in the Fig. 6, a delay time is considered for each task which was selected to be added in a scenario. This means as soon as a scenario executes, a set of selected tasks will be run after its defined time passed from the previous executed task.

User can define each schedule and set a combination of specified schedules as a scenario. Scenarios are used for testing various possible combinations of device states; so different criteria and variables can be simply evaluated, without the need of experimenting on a real environment. The scenarios are designed to set a number of tasks all in one place for further and easier use [11].

IV. EVALUATION

To evaluate the proposed Smart House Simulator we considered a set of scheduled tasks which created by user via the proposed software and then execute them on defined device at the defined time.

In order to create a test plan for simulating via proposed simulator, first user should design a house plane and defines different home appliances and sensors and assign related sensors to each appliance. Then via the "define scheduled task" some tasks should be created on desired objects which are used in the house.

Each task executes on a special device at a specific time and date and set the device's sensor to defined data. Also a scenario can be created via proposed smart house simulator as it described in planning scenarios section.

Scenarios are a set of these tasks which user has create them via "scheduled a task" form without considering their Time and date.

Figure 6: Define Scenario form. A set of scheduled task in the “Scheduled Tasks” checked box, should be selected for defining the scenario. These scheduled are created by user via “Define Schedule” form.

To check the updated status of devices through the simulator, a solution considered that can show status of each device every time the user get the status of them. As it is demonstrated in Fig. 7, by clicking on a specific device and choose “Get Status” option, the updated status of the selective device shown in a message box. So it can be realized if each task has executed correctly or not. Testing a defined scenario is the same also. We only should get the status of all devices which are defined in the scenario as scheduled tasks.

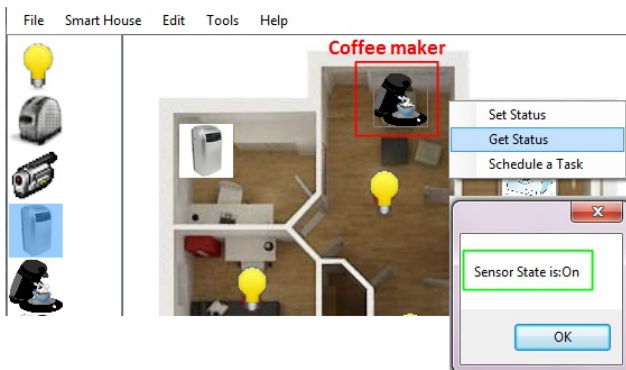


Figure 7: checking the status of selected device

V. CONCLUSIONS

In this paper we presented design and implementation of smart house simulator, which has considered all aspect of smart environment and there is no limitation for using virtual state-of-the-art sensors and appliances for simulation.

The proposed simulator have some characteristics which distinct it from other smart house simulators. The ability of designing the plan of desired house plan or load top view plan of the house as it's plan and then placed appliances in their real place, makes this simulator to simulate a house very close to the real mod.

Scenarios make it easy for people saving the list of actions for further use, in addition to design multiple actions to be done in a single scenario.

The ability of connecting to smart house remote controlling systems is a good choice for researchers in this area to evaluate their projects. Because this feature of propose smart house simulator need a special format for interpreting commands, we considered a unique format for commands which received from house remote controlling systems.

VI. FUTURE WORKS

In the future works, we plan to add the ability to define some variables such as energy consumption, computational complexity and etc. so that we can observe the changes in these variables as the simulation continues. This can be a great leap forward in minimization plans that are about to be applied on real smart houses but the real outcome of plan is hard to estimate as a linear or even nonlinear equation. The simulator can run the house scenarios and calculate real-time values of the variables so we can have a better estimate of the minimization plan if applied to a similar real house.

REFERENCES

- [1] Gamhewage C. de Silva, Toshihiko Yamasaki, and Kiyoharu Aizawa, "An Interactive Multimedia Diary for the Home," *Computer*, vol. 40, no. 5, pp. 52-59, 2007.
- [2] Helal, S., Mann, W., El-Zabadani, H., King, J., Kaddoura, Y., Jansen, E., "The Gator Tech Smart House: a programmable pervasive space," *Computer*, vol.38, no.3, pp. 50-60, 2005.
- [3] Bischoff, Urs, Kortuem, Gerd, "A Compiler for the Smart Space," *Ambient Intelligence European Conference (AMI'07)*, 2007.
- [4] T. Van Nguyen, J. Gook Kim, and D. Choi, "ISS: The Interactive Smart home Simulator," *11th International Conference on Advanced Communication Technology, (ICACT 2009)*, vol.03, pp. 1828-1833, 15-18, 2009.
- [5] O'Neill, E., Klepal, M., Lewis, D. O'Donnell, T., O'Sullivan, D., and Pesch, D., "A testbed for evaluating human interaction with ubiquitous computing environments," *Testbeds and Research Infrastructures for the Development of Networks and Communities Conference*, pp. 60-69, 2005.
- [6] InSu K., HeeMan P., BongNam N., YoungLok L., SeungYong L., and HyungHyo L., "Design and Implementation of Context Awareness Simulation Toolkit for Context learning," *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, vol. 2, pp. 96-103, 2006.
- [7] J. Kaye, D. Castillo, "FlashTM MX for Interactive Simulation," ISBN:14-0181-291-0, THOMSON, 2005.
- [8] Craig Swann, Gregg Caines, "XML in FlashTM", ISBN:0-672-32315-X, QUE Publishing, 2002.
- [9] JoonSeok Park, Mikyeong Moon, Seongjin Hwang, Keunhyuk Yeom, "CASS: A Context-Aware Simulation System for Smart Home", in *Proc. of Fifth International Conference on Software Engineering Research, Management and Applications*, 2007.
- [10] Yerrapragada, C., Fisher, P.S., "Voice Controlled Smart House," *Consumer Electronics, 1993. Digest of Technical Papers. ICCE., IEEE 1993 International Conference on* pp. 154 – 155, 1993.
- [11] Ali Reza Manashty, Amir Rajabzadeh, Zahra Foroootan Jahromi, "A Scenario-Based Mobile Application for Robot-Assisted Smart Digital Homes," *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 8, No. 5, ISSN 1947-5500, Pages 89-96, 2010.

AUTHORS PROFILE

Amir Rajabzadeh received the B.S. degree in telecommunication engineering from Tehran University, Iran, in 1990 and received the M.S. and Ph.D. degrees in computer engineering from Sharif University of Technology, Iran, in 1999 and 2005, respectively. He is currently an assistant professor of Computer Engineering at Razi University, Kermanshah-Iran. He was the Head of Computer Engineering Department (2005–2008) and the Education and Research Director of Engineering Faculty (2008-2010) at Razi University.

Zahra Foroootan Jahromi is a senior B.S. student in Software Computer Engineering at Razi University, Kermanshah. She is now researching in smart environments specially on simulating smart digital homes. Her publications include 4 papers in international journals and conferences and one national conference paper. She has 3 registered national patents and is now teaching Robocop robot designing for elementary and high school students at Alvand guidance school. She is a member of Exceptional Talented Students office of Razi University since 2008 and is a member of The Elite National Foundation of Iran.

Ali Reza Manashty is a M.S student in Artificial Intelligent Computer Engineering at Shahrood University of Thechnology, Shahrood, Iran. He got his B.S. degree in software computer engineering from Razi University ,Kermanshah , Iran, in 2010. He has been researching on mobile application design and smart environments especially smart digital houses since 2009. His publications include 4 papers in international journals and conferences and one national conference paper. He has earned several national and international awards regarding mobile applications developed by him or

under his supervision and registered 4 national patents. He is a member of Exceptional Talented Students office of Razi University since 2008 and is a member of The Elite National Foundation of Iran. He was the teacher assistant of several under-graduate courses since 2008.

Carrier Offset Estimation for MIMO-OFDM Based on CAZAC Sequences

Dina Samaha*, Sherif Kishk, Faye Zaki

Department of electronics and communications engineering
Mansoura University, Egypt

*d_samaha@hotmail.com

Abstract— The combination of Multi-Input Multi-Output (MIMO) with Orthogonal Frequency Division Multiplexing (OFDM) is regarded as a winning technology for future broadband communication. However, its sensitivity to Carrier Frequency Offset (CFO) is a major contributor to the Inter-Carrier Interference (ICI), this effect becomes more severe by the presence of multipath fading in wireless channels. This paper is concerned with CFO estimation for MIMO-OFDM system. The presented algorithm uses a two-step strategy. In the proposed method a preamble structure is used which made up of repeated orthogonal polyphase sequences such as Zadoff and Chu sequences. Both of them belong to the class of Constant Amplitude Zero Auto-Correlation (CAZAC) sequences. The repeated preambles that are constructed using a CAZAC code are simultaneously transmitted from all transmit antennas to accomplish frequency offset estimation. Simulation results show the robustness, accuracy and time-efficiency of the proposed algorithm compared to existing similar algorithms that use PN codes especially in multi-path channel.

Keywords— CFO, MIMO, OFDM, Zadoff-Chu sequences.

I. INTRODUCTION

Nowadays, the limitations of modulation schemes in existing communication systems have become an obstruction in further increasing the data rate. Orthogonal Frequency Division Multiplexing (OFDM) is a promising modulation technique used in a wide range of communications systems. A key aspect of OFDM is the overlapping of individual orthogonal sub-carriers which leads to efficient spectral efficiency. One advantage of OFDM is that it reduces the effect of multipath environments. Multi-Input Multi-Output (MIMO) wireless system is a system that is equipped with multiple antennas at transmitter and receiver, takes spectral efficiency to a new level. MIMO systems are an efficient method to enhance data transmission rate requiring no extra bandwidth in rich scattering environments. The combination of OFDM and MIMO technologies referred to as MIMO-OFDM is a winning combination for wireless technology [1]. MIMO-OFDM has gained more and more interests in recent years. Carrier Frequency Offset (CFO) is caused by the Doppler effect of the channel or the mismatch between the transmitter and receiver. In OFDM systems, CFO destroys the orthogonality between the subcarriers, hence results in Inter Carrier Interference (ICI) and performance degradation. As the core technique is OFDM,

MIMO-OFDM systems are much more sensitive to frequency synchronization errors. Therefore, these errors must be accurately estimated and compensated in order to avoid severe error rates. The synchronization techniques for single-input single-output (SISO) OFDM either exploit the inherent structure of the OFDM symbol using the cyclic prefix part without bandwidth overhead [2]. This approach relies only in the redundancy introduced by the cyclic prefix. Other techniques use specifically designed training symbols consisting of repeated parts [2-4]. Moose [3] proposed a Maximum Likelihood (ML) estimator which can correct the CFO after Fast Fourier Transform (FFT) of two identical training symbols. He also described how to increase the estimation range by using shorter training symbols, on the expense of reducing estimation accuracy. Schmidl and Cox [4] concluded that a first symbol is sent with two identical halves which lead to easier detection based on correlation properties. That is when the CFO is partially corrected in the first training phase, a second training symbol is sent to correct the remaining frequency offset. Tufvesson et al. [5] proposed an approach for frequency offset estimation using Pseudo-Noise (PN) sequence that can correct frequency offset with large estimation range. Recent works tackled the CFO problem in MIMO-OFDM systems [6-9]. Mody and Stuber [6] applied a scheme using orthogonal polyphase sequences as training sequence, to estimate fine frequency offsets in time domain and coarse frequency offsets in frequency domain. Schenk and Van Zelst [7] extended Moose's method using repeated sequence with constant envelope as orthogonal training sequence to realize coarse and fine frequency synchronization in one step in time domain. Yao and Tung-Sang [8] proposed frequency offset estimation in MIMO systems assuming that the frequency offset between the transmit and receive antennas is different, whereas time delay is the same. Recently Liming [9] proposed frequency synchronization scheme using repeated (PN) as training sequences to correct CFO.

In this study an algorithm is proposed based on the idea of [9] which aims to apply its frequency synchronization algorithm for MIMO-OFDM system, using a modified preamble consists of training symbol of constant envelope orthogonal codes with good periodic correlation properties, such as Frank-Zadoff [10] and Chu [11] sequences. Zadoff-

Chu sequences possess good correlation properties which are essential in a variety of engineering applications such as establishing synchronization, performing channel estimation and reducing peak-to-average power ratio. The use of these sequences leads to better frequency offset estimation at each transmitting antenna under the assumption of perfect timing synchronization.

In the proposed method CFO is compensated in two stages, at first CFO correction is performed in an acquisition stage, but there will still be existing residual CFO that has to be compensated. To remove these CFO residuals a tracking stage is implanted.

The paper is organized as follows. Section II describes the system model. In section III frequency synchronization preamble structure is explained in conjunction with properties of Zadoff and Chu sequences. The proposed frequency offset estimation algorithm is explained in section IV. System performance is evaluated through computer simulation in section V. Conclusion remarks are presented in section VI.

II. SYSTEM MODEL

A general MIMO-OFDM system comprising N_t transmitter antennas and N_r receiver antennas is depicted in Fig. 1.

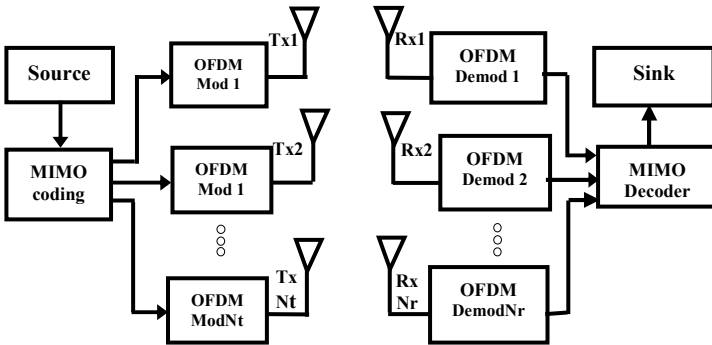


Fig.1. MIMO-OFDM system block diagram.

The received signal on the l^{th} receive antenna is described in equation (1), supposing that time has been synchronized correctly

$$r_l(n) = \sum_{q=1}^{N_t} h_{q,l} s_q(n) \exp\left\{j \frac{2\pi n \epsilon}{L}\right\} + w_l(n) \quad (1)$$

Where $s_q(n)$ is the synchronization training sequence transmitted on q_{th} transmit antenna, $w_l(n)$ is the AWGN on the q^{th} receive antenna, ϵ is the frequency offset factor, L is the length of the training sequence, and $h_{q,l}$ is the channel gain between the q_{th} transmit antenna and the l_{th} receive antenna. In the present study the same frequency offset for each transmit and receive antenna pair has been assumed.

III. SEQUENCE ANALYSIS

The main contribution of the proposed frequency synchronization algorithm is implementation of Zadoff and Chu sequences as synchronization sequences. Both of these sequences are considered as Constant Amplitude Zero Autocorrelation (CAZAC) sequences. The proposed frequency synchronization algorithm uses the good correlation property of CAZAC sequences. It is worthwhile to mention that complexity of polyphase and PN sequences could be compared using two different perspectives. First, a polyphase sequence in frequency domain is also a polyphase sequence in time domain, in other word Zadoff-Chu sequence is also Zadoff-Chu sequence after FFT which can help even in the implementation. Second, the frequency domain correlation of a PN sequence cannot give reliable sequence detection. Thus, keeping in mind the above mentioned comparison it has been decided to choose polyphase sequences as the basic sequence in the proposed study.

A. Frank-Zadoff code sequence

It is defined as cyclic shifted orthogonal code with good periodic correlation properties for preamble sequences. Frank-Zadoff code was described as a more general form of another code introduced by Heimiller [12]. For a code sequence of length L , $\{s_0, s_1, s_2, \dots, s_{L-1}\}$, the complex cyclic correlation function is defined as:

$$x_i = \sum_{n=0}^{L-1} s_{n+i} s_n^* \quad (2)$$

note that s^* denotes the complex conjugate of s , $s_m = s_{m+L}$ because it is a cyclic code. For $i = 0$, the value of x_i reaches its maximum:

$$x_0 = \sum_{n=0}^{L-1} |s_n|^2 \quad (3)$$

On the other hand, for $0 < i \leq L-1$ the values of x_i should be zero. That means each code is orthogonal to its own phase shifted version.

B. Chu sequence

The autocorrelation function of Chu sequences is known to be zero except at the lag of an integer multiple of the sequence length. The length of Zadoff codes is restricted to perfect squares. But, Chu sequences have the same correlation properties and can be constructed for any code length. A set of Chu sequence with length L is considered as S_n , $0 < n < L-1$, where the k_{th} element of S_n , $S_n(k)$, is defined as:

$$S_n(k) = \begin{cases} \exp(j\pi \frac{nk^2}{L}) & ; L \text{ even} \\ \exp(j\pi \frac{nk(k+1)}{L}) & ; L \text{ odd} \end{cases} \quad (4)$$

C. Preamble design

The data packet is preceded by a section of pre-defined data, preamble, which is constructed using a repeated CAZAC code. Each transmitter transmits the same code, but with different cyclic shifts. The preamble is followed by a data transmission on all transmitters. The advantage of this MIMO preamble is that it takes only as much time as a SISO preamble and it is independent of the number of transmitting antennas. For the proposed frequency synchronization algorithm a repetition of the CAZAC training sequence is considered. The training sequences from different transmit antennas have to be orthogonal to each other for at least the maximum channel delay length to preserve orthogonality.

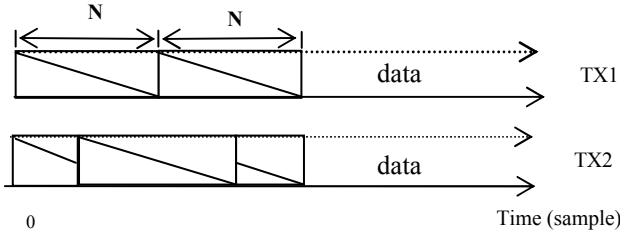


Fig.2. MIMO-OFDM system preamble schematic diagram using 2 transmitting antennas

Fig. 2 shows an example of preamble including a CAZAC sequence repetition with N periods for 2 transmitters MIMO system. It is transmitted twice by 2 transmitters simultaneously with a different cyclic shift.

IV. FREQUENCY SYNCHRONIZATION ALGORITHM

The proposed CFO estimation specifies a unique training sequence at each transmit antenna to designate the antennas and estimate the CFO. Chu and Zadoff sequences are adopted as training sequences in different transmit antenna with their shift and orthogonal properties. Assume the length of training sequence is L , and the period length of Chu or Zadoff sequence is N , then $M=L/N$. M is positive integer pointed to number of Chu or Zadoff sequences contained in the training sequence. Generally CFO estimation will be switched between two operation modes, the first called acquisition mode and the other is tracking mode. In the acquisition mode, a wide range of CFO can be estimated and the remaining CFO should be much less than half of the space between subcarriers. During the tracking mode only small frequency fluctuations will be dealt with

A. CFO Acquisition

CFO acquisition is performed only once when the transmissions begin. Therefore, the time of acquisition is not critical. When there is no channel fading and noise, the relationship between corresponding samples from different Chu or Zadoff sequences in a received training sequence on the same antenna is given by:

$$r_l(n) \exp\left(\frac{j2\pi g \varepsilon L}{N}\right) \quad (5)$$

$$l \in (1, Nr), n \in (0, L - gN - 1)$$

Where g is the correlation period (the number of Chu or Zadoff sequences) denoting the distance between two correlated samples. The CFO can be estimated as:

$$\varepsilon = \frac{L}{2\pi g N} \times \text{angle}(\phi_g), \quad g \in (1, M-1) \quad (6)$$

Where ϕ_g is defined as:

$$\phi_g = \sum_{l=1}^{N_t} \sum_{n=0}^{L-gN-1} r_l(n + gN) r_l^*(n) \quad (7)$$

In (6), ε is estimated with single “ g ”. This kind of estimators is named as single- g estimators. They are adopted for CFO acquisition, the estimation range is $|\varepsilon| < \frac{L}{2gN}$.

B. CFO Tracking

For different “ g ”, multiple different single- g estimators can be used together in estimating CFO to improve estimation accuracy. This is called as multiple- g estimator. The number of used “ g ” will be pointed to by parameter “ m ”. One multiple- g estimator is given by:

$$\varepsilon = \sum_{j=1}^m \frac{L \times \text{angle}(\phi_{g_j}) \times (2\pi g_j N)^{-1}}{m} \quad (8)$$

Where: $m \in (1, M-1)$, $g_1, g_2, g_3, \dots, g_m \in (1, M-1)$

The estimation range of multiple- g estimator is $|\varepsilon| < \frac{L}{2N \max(g_1, g_2, \dots, g_m)}$. The estimation error of one

multiple- g estimator may be much smaller than each single- g estimator used by it, especially when all these single- g estimators have the same or close accuracy.

V. SIMULATION RESULTS

In order to examine the synchronization performance, number of simulation experiments will be performed for the proposed synchronization scheme and that described in [9]. Results reported here are carried out using Matlab[®]. There are a number of parameters that describe the system. In the simulations, the following parameters are held constant:

- (1) Length of the training sequences: $L = 1024$.
- (2) Normalized CFO factor=0.5, that uniformly distributed within ± 0.5 subcarrier spacing.
- (3) The multipath channel consists of six paths that have uniformly distributed delays over the interval $[0, 2\pi]$.
- (4) Results are calculated for a 6x6 MIMO-OFDM system under the assumption of ideal modulation.

The performance of the estimator is evaluated by the Mean-Square Error (MSE) of the frequency offset estimates. The MSE for the p -th transmit antenna is defined as

$$MSE_p = \frac{1}{N_{matlab}} \sum_{i=1}^{N_{matlab}} |\varepsilon_{est,p} - \varepsilon_p|^2 \quad (9)$$

Where $\varepsilon_{est,p}(i)$ the frequency is offset estimate obtained in the i -th Matlab[®] trial, and N_{matlab} is the total number of Matlab[®] trials.

In the first simulation experiment, comparison results of using PN and zaddoff sequences in AWGN channel are shown in Fig.(3). Single-g estimator ($g=1$) (acquisition) and multiple-g estimators ($m=2, g=2, 3$) (tracking) are considered for both used sequences. The using preamble consists of four repeated sequences and, $N=256$ (the period length of training sequence). It can be shown that in acquisition stage performance of zaddoff sequence better than PN sequence. In the tracking mode it is obvious that using zaddoff sequence improves the estimation of CFO.

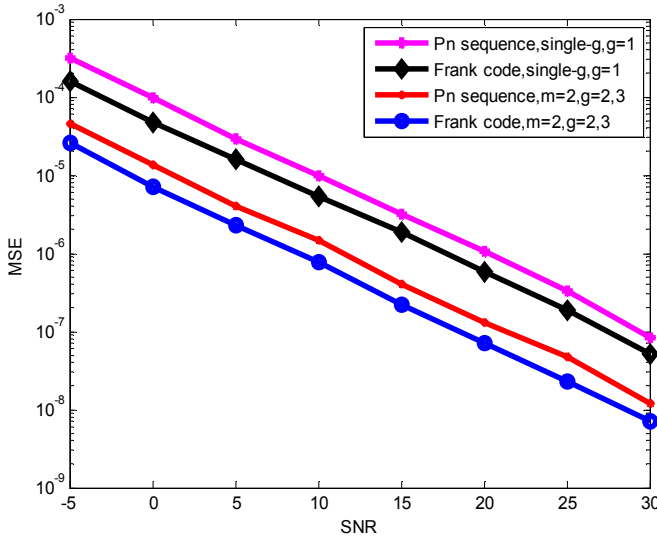


Fig.3. MSE versus SNR for PN and zaddoff sequence in acquisition And tracking in AWGN ch.

To obtain the results of Fig.(4), simulation experiment results run using Chu sequences compared with PN sequences in AWGN channel. The using preamble structured of 8 repeated sequences ($N=128$). Single-g estimator ($g=2$) (acquisition) and multiple-g estimators ($m=4, g=3, 4, 5, 6$) (tracking) are considered for both PN and Chu sequences. The comparison obviously illustrates that Chu sequence performs much better in acquisition stage and it is more accurate for low and high different value of “ g ” in tracking stage.

From Figs.3 and 4, it can be seen that the performance of zaddoff and Chu sequences in multiple-g estimator is better than that of single-g estimator. Both sequences impact on the offset estimation performance nearly the same, which gives wide choices to enhance the system performance.

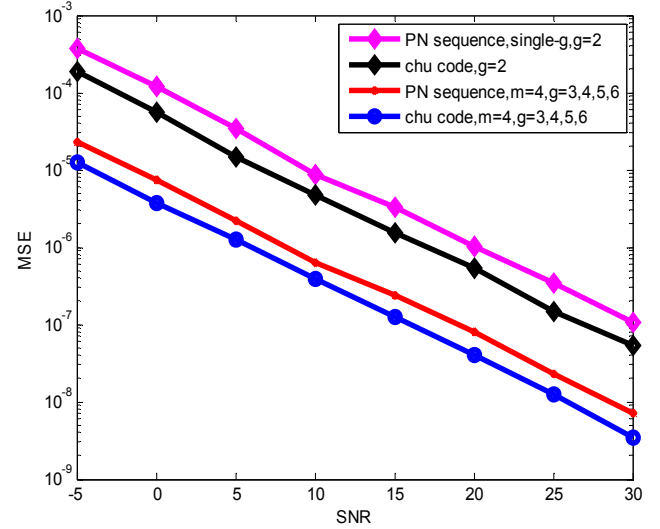


Fig.4. MSE versus SNR for PN and Chu sequence in acquisition and tracking in AWGN ch.

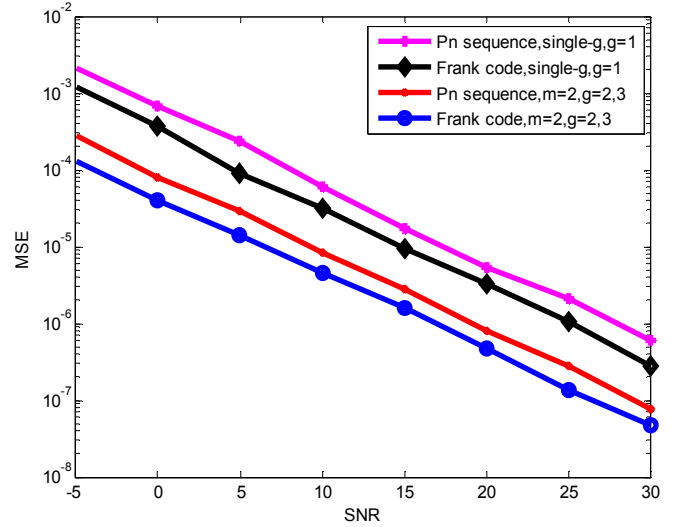


Fig. 5 MSE versus SNR for PN and zaddoff sequence in acquisition and tracking in multipath ch.

The other side in our simulation experiment is examining the sequences influence in the estimation with the effect of the multipath channel described before. In Fig.(5) and Fig.(6) the same parameters mentioned before are assumed for simulating, but with the effect of multipath channel. The better performance comparison of zaddoff code and Chu code with PN sequence is demonstrated.

From all previous simulation experiments, it can be seen that the performance of zaddoff and Chu sequences is identical; they are nearly with the same trend, with same variations in all the simulation experiments depicted previously.

In general both sequences impact on the offset estimation performance nearly the same, which gives wide choices to enhance the system performance, the proposed preamble using either zaddoff sequence or chu sequence as training sequence is more robust and accurate from two sides. The first one is its superior performance and the other is saving required time

which makes the system more accurate especially with existing of the multipath effects

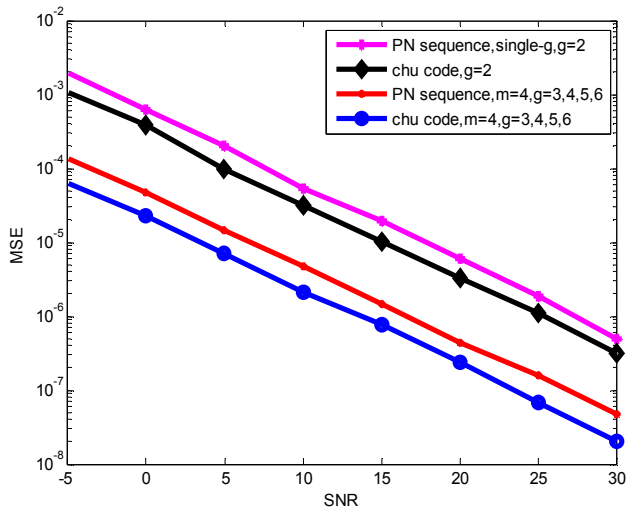


Fig.6 MSE versus SNR for PN and Chu sequence in acquisition and tracking in multipath ch

VI. CONCLUSION

A new CFO synchronization method for MIMO-OFDM systems with modified preamble method using repeating CAZAC sequences is carried out. The performance of a two stage synchronization structure has been studied. Simulation results show that fast and robust synchronization can be established. The proposed method enhances the synchronization performance even under low SNR. The new method surpasses the traditional one using regular PN sequences in both the AWGN channel and multipath channel. It achieves better performance without consuming much computation time. This is considered a very important feature in wireless communication systems in general.

REFERENCES

[1] WILLINK T.J.: 'MIMO OFDM for broadband fixed wireless access', IEEE Proc. Commun., 2005, 152, (1), pp. 75-81
[2] J. Van de Beek, M. Sandell, and P. O. Brjesson, "ML Estimation of

[3] Time and Frequency Offset in OFDM Systems," IEEE Trans. Signal Processing, vol. 45, pp. 1800-1805, July 1997.
[4] P. H. Moose, "A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction," IEEE Trans. Commun., vol. 42, pp. 2908-2914, October 1994.
[5] T. M. Schmidl and D. C. Cox, "Robust Frequency and Timing Synchronization for OFDM," IEEE Trans. Commun., vol. 45, pp. 1613-1621, December 1997.
[6] F. Tufvesson, M. Faulkner and O. Edfors, "Time and frequency synchronization for OFDM using PN-sequence preambles," Proceedings of IEEE Vehicular Technology Conference, Amsterdam, September 19-22, pp. 2203-2207, 1999.
[7] A. N. Mody and G. L. Stuber, "Receiver Implementation for a MIMO OFDM System," Proc. of IEEE GLOBECOM 2002, vol. 1, pp. 716-720, November 2002.
[8] T. C. W. Schenk and A. Van Zelst, "Frequency Synchronization for MIMO OFDM Wireless LAN Systems," Proc. of IEEE VTC 2003-Fall, vol. 2, pp. 781-785, October 2003.
[9] Y. Yao, and Tung-Sang Ng, "Correlated-Based Frequency offset Estimation in MIMO System", Proc. IEEE Vehicular Technology Conference Fall 2003 (VTC Fall 2003), Orlando (FL), 6-9 October 2003.
[10] LimingHe, "Frequency Synchronization in MIMO OFDM Systems," Wireless Communications Networking and Mobile Computing Conference (WICOM), China, 2010.
[11] R. Frank, and S Zadoff, "Phase Shift Pulse Codes With Good Periodic Correlation Properties," IRE Trans. on Inform. Theory, IT-8, 381-382, 1962.
[12] D. C. Chu, "Polyphase Codes With Good Periodic Correlation Properties," IEEE Trans. Info. Theory 18,531-532 (July 1972).
[13] R. C. Heimiller, "Phase Shift Pulse Codes With Good Periodic Correlation Properties," IRE Trans. Info. Theory IT-6, 254- 257 (October 1961).

A New Approach to Prevent Black Hole Attack in AODV

M. R. Khalili Shoja

Department of Electrical
Engineering, Amirkabir University
of Technology, Tehran, Iran
m.khalili@aut.ac.ir

Hasan Taheri

Department of Electrical
Engineering, Amirkabir University
of Technology, Tehran, Iran
htaheri@aut.ac.ir

Shahin Vakiliinia

Department of Electrical
Engineering, Sharif University
of Technology, Tehran, Iran
vakiliinia@ee.sharif.edu

Abstract— Ad-hoc networks are a collection of mobile hosts that communicate with each other without any infrastructure. These networks are vulnerable against many types of attacks including black hole. In this paper, we analyze the effect of this attack on the performance of ad-hoc networks using AODV as a routing protocol. Furthermore, we propose an approach based on hash chain to prevent this type of attack. Simulation results using OPNET simulator depicts that packet delivery ratio, in the presence of attacker nodes, reduces remarkably. On the other hand, applying proposed approach can reduce the effect of black hole attacks.

Keywords: AODV; black hole; hash chain; OPNET

I. INTRODUCTION

Ad-hoc networks are characterized by dynamic topology, self-configuration, self-organization, restricted power, temporary network and lack of infrastructure. Characteristics of these networks lead to using them in disaster recovery operation, smart buildings and military battlefields [1].

Routing protocol in ad-hoc networks are classified into two main categories, proactive and reactive [3]. In proactive routing protocols, routing information of nodes is exchanged, periodically, such as DSDV [4]. In on-demand routing protocols, nodes exchange routing information when needed such as, AODV [2] and DSR [5]. Furthermore, some ad-hoc routing protocols are a combination of above categories.

Although trusted environment has been assumed in most research on ad-hoc routing protocols, many usages of ad-hoc network run in untrusted situations. So, most ad-hoc routing protocols are vulnerable to diverse types of attacks that one of which is black hole attack. In this attack, a malicious node uses the routing protocol to advertise itself as having the shortest or freshest path to the node whose packets it wants to intercept. In a flooding based protocol, the attacker listens to requests for routes. When the attacker receives a request for a route to the target node, the attacker creates a reply consisting of an extremely short or fresh route [6]. The rest of this paper is organized as follows: In section 2, AODV routing protocol is described. In section 3, we describe classification of attacks in MANET. Network layer attack is described in section 4. Section 5 summarizes related works and detailed description of

proposed solution is discussed in section 6. In section 7, simulation results are analyzed.

II. AODV ROUTING PROTOCOL

AODV is used to find a route between source and destination as needed and this routing protocol uses three significant type of messages, route request (RREQ), route reply (RREP) and route error (RERR). Field information of these messages, such as source sequence number, destination sequence number, hop count and etc is explained in detail in [2]. Each node has a routing table, which contains information about the route to the specific destination. When source node wants to communicate with a destination and there is not any route between them in its routing table, at first step source node broadcasts RREQ. So, RREQ is received by intermediate nodes that they are in the transmission range of sender. These nodes broadcast RREQ until RREQ is received by destination or an intermediate node that has fresh enough route to the destination. Then it sends RREP unicastly toward the source. Hence, a route among source and destination is made. A fresh enough route is a valid route entry that its destination sequence number is at least as great as destination sequence number in RREQ. The source sequence number is used to determine freshness about route to the source. In addition, destination sequence number is used to determine freshness of a route to the destination. When intermediate nodes receive RREQ, with consideration of source sequence number and hop count, make or update a reverse route entry in its routing table for that source. Furthermore, when intermediate nodes receive RREP, with consideration of destination sequence number and hop count, make or update a forward route entry in its routing table for that destination.

III. CLASSIFICATION OF ATTACKS IN MANET

The attacks in MANET can be classified into two categories, called passive attacks and active attacks. Passive attacks are done to steal information of network such as, eavesdropping attacks and traffic analysis attacks. Indeed, passive attackers get data exchanged in the network without disrupting the operation of a network and modification of exchanged data. On the other hand, in active attacks, replication, modification and deletion of exchanged data is

done by attackers. The attacks in ad-hoc networks can also be classified into two categories, called external attacks and internal attacks. Internal attacks are done by authorized node in the network, whereas external attacks are executed by nodes that they are not authorized to participate in the network options. Another classification of attacks is related to protocol stacks, for instance, network layer attacks.

IV. NETWORK LAYER ATTACKS IN MANET

Some network layer attacks are described in below:

A. Wormhole attack

In this attack, an attacker records a packet, at one location in the network, tunnels the packet to another location and replays it there [21].

B. Byzantine attack

In this attack, malicious nodes individually or cooperatively carry out attacks such as creating routing loops and forwarding packets through non-optimal paths.

C. Rushing attack

Rushing attacker forwards packets quickly by skipping some of the routing processes. So, in on-demand routing protocol such as AODV, the route between source and destination include rushing nodes.

D. Resource consumption attack

In this attack, an attacker attempts to consume battery life of other nodes.

E. Location disclosure attack

In this attack, information relating to structure of network is revealed by attacker nodes.

F. Black hole attack

In black hole attack, malicious nodes falsely claim a fresh route to the destination to absorb transmitted data from source to that destination and drop them instead of forwarding.

Black hole attack in AODV protocol can be classified into two categories: black hole attack caused by RREP and black hole attack caused by RREQ.

1) Black hole attack caused by RREQ

With sending fake RREQ messages an attacker can form black hole attack as follows:

- Set the originator IP address in RREQ to the originating node's IP address.
- Set the destination IP address in RREQ to the destination node's IP address.
- Set the source IP address of IP header to its own IP address.
- Set the destination IP address of IP header to broadcast address.
- Choose high sequence number and low hop count and put them in related fields in RREQ.

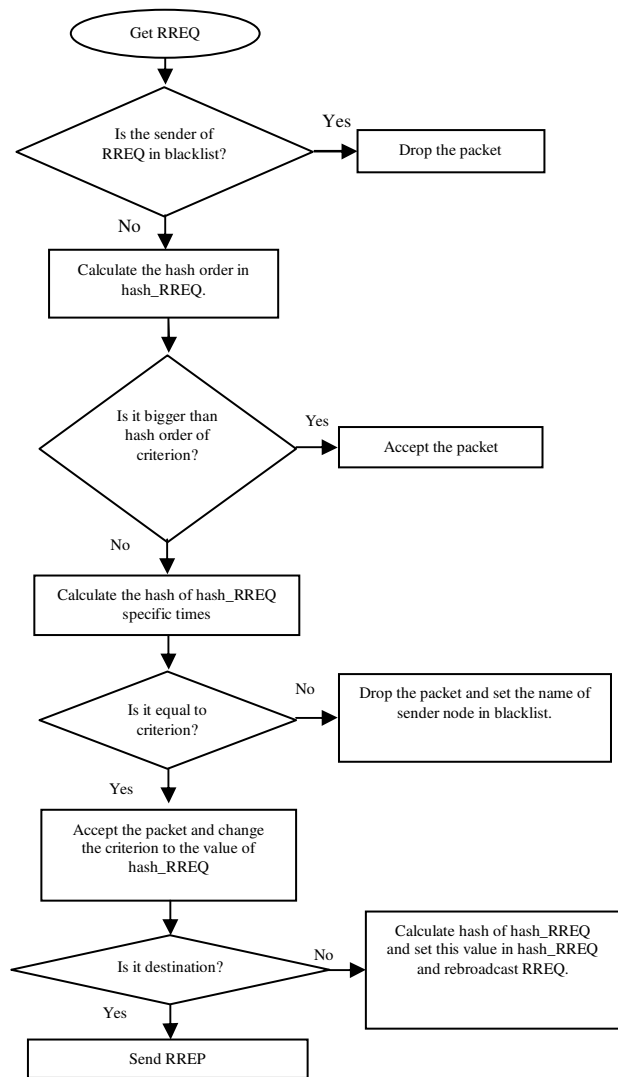


Figure 1. Operation at intermediate nodes when receive RREQ

So, false information about source node is inserted to the routing table of nodes that get fake RREQ. Hence, if these nodes want to send data to the source, at first step they send it to the malicious node.

2) Black hole attack caused by RREP

With sending fake RREP messages an attacker can form black hole attack. After receiving RREQ from source node, a malicious node can generate black hole attack by sending RREP as follow:

- Set the originator IP address in RREP to the originating node's IP address.
- Set the destination IP address in RREP to the destination node's IP address.
- Set the source IP address of IP header to its own IP address.
- Set the destination IP address of IP header to the IP address of node that RREQ has been received from it.

TABLE I. SIMULATION PARAMETERS

Simulation parameters	Value
Number of nodes	46
Network size	600*600(m)
Simulation duration	600(sec)
Transmit power(w)	.0001
Packet Reception-power Threshold(dBm)	-95
Hash function	SHA-1
Source node	Mobile-node-1
Destination node	Mobile-node-4
Packet Inter-Arrival Time(sec)	Uniform(.1,.11)
Packet size(bits)	Exponential(1024)

e) Choose high number for sequence number and low number for hop count.

So, data from source reach to malicious node and it drops them.

V. RELATED WORKS

There are basically two approaches to secure MANET: 1. securing ad-hoc routing and 2. Intrusion detection [7].

A. Securing routing

Ariadne [8] has proposed ad-hoc routing protocol that provides security in MANET and relies on efficient symmetric cryptography. This protocol is based on the basic operation of the DSR protocol. In [9], a secure routing protocol based on DSDV has been proposed. Hash chains have been used to authenticate hop counts and sequence numbers. ARAN [10] uses cryptographic public-key certificates in order to achieve the security goals. The goal of SAR [11] is to characterize and explicitly represent the trust values and trust relationships associated with ad-hoc nodes and use these values to make routing decisions. Secure AODV (SAODV) [12] is a security extension of AODV protocol, based on public key cryptography. Hash chains are used in this protocol to authenticate the hop count. Adaptive SAODV (A-SAODV) [13] has proposed a mechanism based on SAODV for improving the performance of SAODV. In [14] a bit of modification has been applied to A-SAODV for increasing its performance. TRP [20] employs hash chain algorithm to generate a token, which is appended to the data packets to identify the authenticity of the routing packets and to choose correct route for data packets. TRP provides significant reduction in energy consumption and routing packet delay by using hash algorithm.

B. Intrusion detection system

[15] introduces a method that requires each intermediate node to send back the next hop information inside RREP message. This method uses further request message and further reply

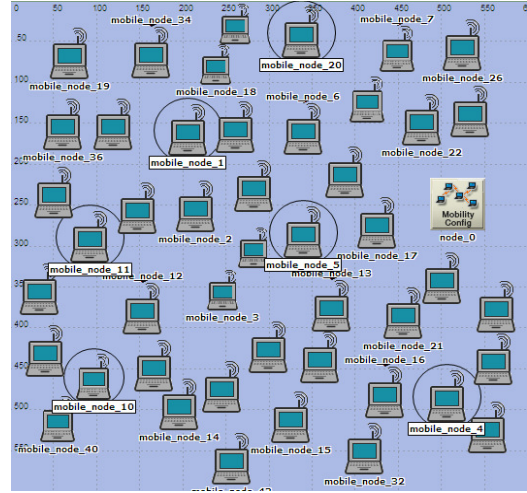


Figure 2. Network topology

message to verify the validity of the route. Zhang and Lee [16] propose a distributed and cooperative intrusion detection model based on statistical anomaly detection techniques. In [17], the intermediate node requests its next hop to send a confirmation message to the source. After receiving both route reply and confirmation message, the source determines the validity of path according to its policy. In [18], Huang et al use both specification-based and statistical-based approaches. They construct an Extended Finite State Automation (EFSA) according to the specification of AODV routing protocol and model normal state and detect attacks with anomaly detection and specification-based detection. An approach based on dynamic training method in which the training data is updated at regular time intervals has been proposed in [19].

VI. PROPOSED WORK

As we will discuss, AODV is weak against black hole attack. In this paper we propose mechanism to prevent black hole attack based on hash chain. Black hole attack is based on modification of sequence number and hop count. In this method, when an intermediate node receives RREQ or RREP, check an extra field, which will be explained later, to verify sequence number and hop count. If this node authenticates validity of this field, it can accept control messages and fill its routing table accordingly. We add one field named hash_RREQ to RREQ and similarly, one field called hash_RREP to RREP. We assume that only destination can send RREP, although, intermediate nodes can have enough fresh routes to the destination. It means that destination only flag in RREQ must be set. When source node wants to send data to destination, it must use AODV protocol to find a route to the destination. So this node sends new RREQ as below:

Normal RREQ	Hash-RREQ
-------------	-----------

Normal RREQ is RREQ in AODV and hash_RREQ field is a new field that we add to existing protocol. Source node

should fill this field and intermediate nodes should verify the authenticity and change this field as will be explained. Source node chooses a random number as a seed. After that this node calculates hash of the seed a specific number of times as below:

$$h_{(a-i),b+j}(seed) \quad (1)$$

Where $h_r(x)$ means calculation of hash of x , r times (or order of hash is r) and a is the number that should be higher than maximum sequence number during the working of network, b is the maximum hop count between two furthest nodes plus one in the network, i is the value of sequence number and j is the hop count (patently j for source node is 0). We assume that before sending the RREQ with source node, all nodes in the network know the value:

$$h_{(a+1),b}(seed) \quad (2)$$

We called this value criterion. Therefore, source node generates the hash of the seed using the formula (1) and places it in the hash_RREQ field; for example, for first RREQ that is sent by source node, $i=1$ and $j=0$. When this packet, RREQ, arrives to next hop, at the first step, this node checks the validity of hash value by knowing the formula (2). It means that the receiver node computes hash of hash_RREQ for y times, where y is the difference between order of hash in hash_RREQ and order of hash in criterion. If it is equal to criterion, Eq. (2), so intermediate node accepts this packet and inserts appropriate entry in relation with source node in its routing table. On the other hand, receiver node, after validating of hash_RREQ, changes its criterion to the value of hash_RREQ. Now this node should change the hash_RREQ in accordance with 1. It means that this node calculates hash of the hash_RREQ and places it in hash_RREQ.

Similarly, destination node does as above but calculates hash chain in accordance with its own seed. Destination node places the value of hash in hash_RREP. New RREP is shown below:

Normal RREP	Hash-RREP
-------------	-----------

Other things and operation are exactly similar to RREQ. Thus, malicious nodes cannot increase sequence number and decrease hop count. Because if they want do this, they should calculate hash in lower order and obviously this is impossible, for instance, with having h_{20} it is impossible to calculate h_{19} . Consequently, malicious nodes can only increase hop count or decrease sequence number. As a result, wrong information about sequence number and hop count cannot be placed in routing tables. Besides, each node has a blacklist and when a node receive RREQ or RREP from malicious node that hash_RREQ or hash_RREP field is not valid, puts the name of the sender in its own blacklist. Furthermore all packets from nodes in blacklist won't be accepted and must be discarded. Operation at intermediate nodes, when receive RREQ, is

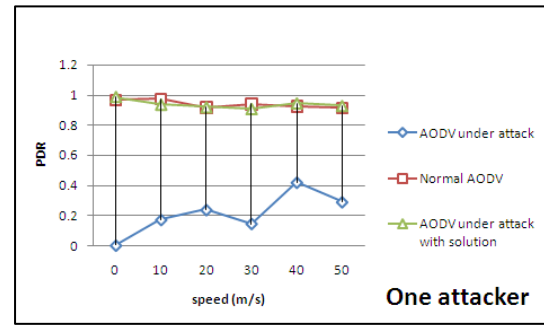


Figure 3. PDR in presence of one malicious node



Figure 4. PDR in presence of two malicious nodes

depicted in Fig. 1. Furthermore, operation at intermediate nodes, on receiving RREP, is exactly the same as Fig. 1.

VII. SIMULATION RESULTS

For the simulation, we use OPNET 14.0.A [22] as a simulator. Our network topology is indicated in Fig. 2. TABLE I contains parameters that we choose for simulation. For evaluating the performance of the network, we consider the following metrics:

Packet Delivery Ratio: The ratio of the data delivered to the destination to the data sent out by the source.

Various mobilities of nodes have been considered to measure the performance of network in presence of malicious nodes as attackers. Fig. 3 demonstrates the results in presence of only one malicious node. In this scenario mobile-node-5 is the attacker. In Fig. 4, mobile-node-5 and mobile-node-20 act as malicious nodes. Mobile-node-5, mobile-node-20 and mobile-node-10 are malicious nodes in another scenario and results of this section are presented in Fig. 5. In another scenario mobile-node-5, mobile-node-20, mobile-node-10 and mobile-node-11 send fake messages to construct a black hole attack. Related results are illustrated in Fig. 6. Fig. 7, Fig. 8 and Fig. 9 show the variation of PDR when the number of malicious nodes in the network is varied from 1 to 4 in the mobility of 10m/s, 30m/s, 50m/s respectively. As it is shown by these figures, the proposed mechanism improves PDR nodes, in other words, this approach protects MANETs against black hole attack.

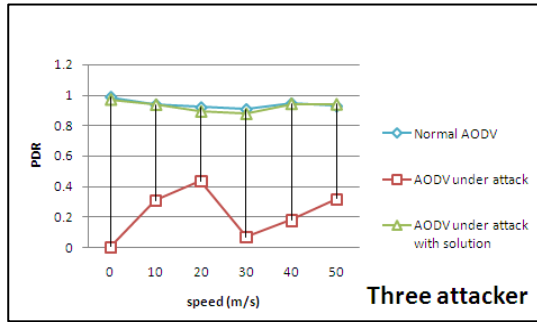


Figure 5. PDR in presence of three malicious nodes

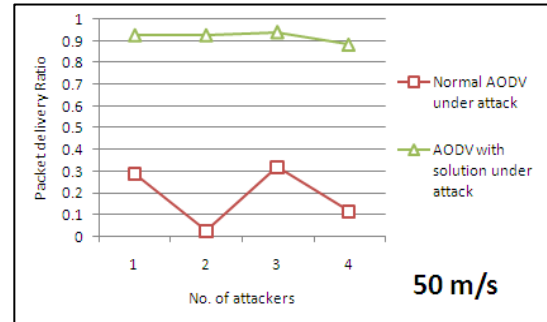


Figure 9. Analysis of PDR vs. different number of attacker in 50m/s

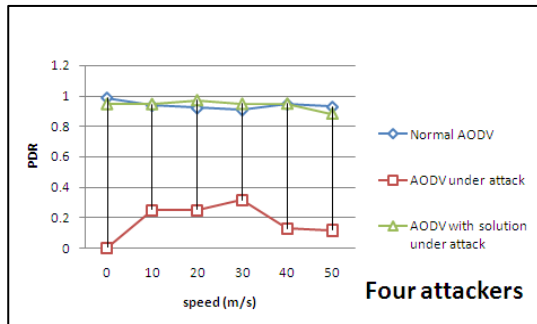


Figure 6. PDR in presence of four malicious nodes

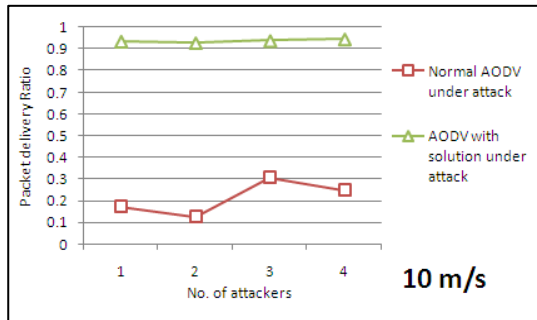


Figure 7. Analysis of PDR vs. different number of attacker in 10m/s

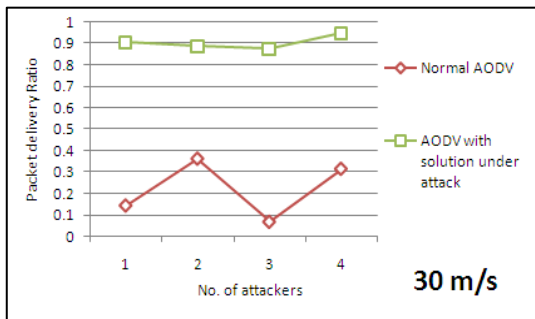


Figure 8. Analysis of PDR vs. different number of attacker in 30m/s

VIII. REFERENCES

- [1] E. Çayırıcı, C.Rong, "Security in Wireless Ad Hoc and Sensor Networks," vol. I. New York: Wiley 2009, pp. 10.
- [2] C. E. Perkins, E. M. B. Royer and S. R. Das, "Ad-hoc On-Demand Distance Vector (AODV) Routing," Mobile Ad-hoc Networking Working Group, Internet Draft, draft-ietf-manetaadv-00.txt, Feb. 2003.
- [3] E. M. Royer and C-K Toh, "A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks," IEEE Person. Commun., Vol. 6, no. 2, Apr. 1999.
- [4] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," Proceedings of the SIGCOMM '94 Conference on Communications Architectures, Protocols and Applications, pp 234-244, Aug.1994.
- [5] D. B. Johnson, D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks", Mobile Computing, edited by Tomasz Imielinski and Hank Korth, Chapter 5, pp 153- 181, Kluwer Academic Publishers, 1996.
- [6] M. Ilyas, "The Handbook of Ad hoc wireless Networks," CRC Press, 2003.
- [7] Stamouli, "Real-time Intrusion Detection for Ad hoc Networks" Master's thesis, University of Dublin, Sep. 2003.
- [8] Y.-C. Hu, A. Perrig, D. B. Johnson, "Ariadne: A Secure On-Demand Routing Protocol for Ad hoc Networks," Proc. 8th ACM Int'l. Conf. Mobile Computing and Networking (Mobicom'02), Atlanta, Georgia, Sep. 2002, pp. 12-23.
- [9] Y.-C. Hu, D. B. Johnson, A. Perrig, "SEAD: Secure Efficient Distance Vector Routing for Mobile Wireless Ad hoc Networks," Proc. 4th IEEE Workshop on Mobile Computing Systems and Applications, Callicoon, NY, Jun. 2002, pp. 3-13.
- [10] K. Sanzgiri, B. Dahill, B. Levine, C. Shields, and E. Belding-Royer, A Secure Routing Protocol for Ad Hoc Networks. Proc. of IEEE International Conference on Network Protocols (ICNP), pp. 78-87, 2002.
- [11] S. Yi, P. Naldurg, R. Kravets, "Security-Aware Ad hoc Routing for Wireless Networks," Proc. 2nd ACM Symp. Mobile Ad hoc Networking and Computing (MobiHoc'01), Long Beach, CA, Oct. 2001, pp. 299-302.
- [12] M. Zapata, "Secure Ad Hoc On-Demand Distance Vector (SAODV)," Internet draft, draft-guerrero-manet-saodv-01.txt, 2002.
- [13] D. Cerri, A. Ghioni, "SecuringAODV: The A-SAODV Secure Routing Prototype," IEEE Communication Magazine, Feb. 2008, pp 120-125.
- [14] K. Mishra, B. D. Sahoo, "A Modified Adaptive-Saodv Prototype For Performance Enhancement In Manet," International Journal Of Computer Applications In Engineering, Technology And Sciences (Ij-Ca-Ets), Apr. 2009 – Sep. 2009, pp 443-447.

- [15] H. Deng, W. Li, and D. P. Agrawal, "Routing security in ad hoc networks," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 70-75, Oct. 2002
- [16] Y. Zhang and W. Lee, "Intrusion detection in wireless ad – hoc networks," 6th annual international Mobile computing and networking Conference Proceedings, 2000.
- [17] S. Lee, B. Han, and M. Shin, "Robust routing in wireless ad hoc networks," in *ICPP Workshops*, pp. 73, 2002.
- [18] Y. A. Huang and W. Lee, "Attack analysis and detection for ad hoc routing protocols," in *The 7th International Symposium on Recent Advances in Intrusion Detection (RAID'04)*, pp. 125-145, French Riviera, Sept. 2004.
- [19] S. Kurosawa, H. Nakayama, N. Kat, A. Jamalipour, and Yoshiaki Nemoto, "Detecting Blackhole Attack on AODV-based Mobile Ad Hoc Networks by Dynamic Learning Method", *International Journal of Network Security*, Vol.5, No.3, pp 338-346, Nov. 2007 .
- [20] L. Li , C. Chigan, "Token Routing: A Power Efficient Method for Securing AODV Routing Protocol," *Proceedings of the 2006 IEEE International Conference on Networking, Sensing and Control*, 2006. ICNSC '06, (Apr. 2006) pp 29- 34.
- [21] Y. C. Hu, A. Perrig, D. B. Johnson, "Wormhole Attacks in Wireless Networks," *Ieee Journal On Selected Areas In Communications*, Vol. 24, No. 2, Feb. 2006.
- [22] <http://www.opnet.com>

Application of Web Server Benchmark using Erlang/OTP R11 and Linux

A. Suhendra¹, A.B. Mutiara²

Faculty of Computer Science and Information Technology, Gunadarma University
Jl. Margonda Raya No.100, Depok 16424, Indonesia
^{1,2}{adang, amutiara}@staff.gunadarma.ac.id

Abstract— As the web grows and the amount of traffics on the web server increase, problems related to performance begin to appear. Some of the problems, such as the number of users that can access the server simultaneously, the number of requests that can be handled by the server per second (requests per second) to bandwidth consumption and hardware utilization like memories and CPU. To give better quality of service (QoS), web hosting providers and also the system administrators and network administrators who manage the server need a benchmark application to measure the capabilities of their servers. Later, the application intends to work under Linux/Unix – like platforms and built using Erlang/OTP R11 as a concurrent oriented language under Fedora Core Linux 5.0. It is divided into two main parts, the controller section and the launcher section. Controller is the core of the application. It has several duties, such as read the benchmark scenario file, configure the program based on the scenario, initialize the launcher section, gather the benchmark results from local and remote Erlang node where the launcher runs and write them in a log file (later the log file will be used to generate a report page for the sysadmin). Controller also has function as a timer which act as timing for user inters arrival to the server. Launcher generates a number of users based on the scenario, initialize them and start the benchmark by sending requests to the web server. The clients also gather the benchmark result and send them to the controller.

Key words— Erlang, QoS, Network Management, Concurrent Programming, Distribution

I. INTRODUCTION

In the last two decades, human necessities in fast and accurate information create a lot of innovations in information technology, one of them is the internet. Since TCP/IP released to public in 1982 and World Wide Web (WWW) introduced in 1991, internet has become a popular media to access and publish information. The easy to use web mechanisms make people easy to search and publish information on the internet. The web service later grows to many aspects, such as entertainment, education, scientific research and many more.

To access the web on the internet, we need a certain server than can provide user access on the web pages. This server is called web server or HTTP server and has a main duty to serve user access to web pages contents, either static or dynamic.

To give better quality of service (QoS) , web hosting providers and also the system administrators and network administrators who manage the server need a benchmark application to measure the capabilities of their servers. Later,

the application intends to work under Linux/Unix -- like platforms and built using Erlang/OTP R11 as a concurrent oriented language under Fedora Core Linux 5.0.

Based on the above descriptions, there are some problems than can be summarized, such as:

- 1) To give better Quality of Service, web hosting provider and also system administrators and network administrators who manage the web server need a benchmark application to measure the capabilities/performances of their servers.
- 2) The benchmark application is intended to be use by the network administrators and system administrators who work under Linux/Unix – like systems.
- 3) The application is made by utilizing the concurrent capability of Erlang programming language under Linux operating system.

II. THEORIES

A. Web Server

The term web server can mean one of two things [2]:

- 1) A computer or a number of computers which responsible for accepting HTTP requests from clients, which are known as web browsers, and serving them web pages, either static or dynamic pages.
- 2) A computer program that provides the functionality described in the first sense of the term.

Web server also works based on several standards, such as [2]: HTTP response to HTTP Request, Logging, Configurability, Authentication, Handling Static and Dynamic Contents, Modular Support, Virtual Hosts

B. Erlang/OTP

Erlang is a concurrent programming language with a functional core. By this we mean that the most important property of the language is that it is concurrent and that secondly, the sequential part of the language is a functional programming language. Concurrent means that the language has focus on how to makes multiple executions threads to run and do computational work together. In Erlang, these execution threads are called processes. The sequential sub-set of the language expresses what happens from the point it time

where a process receives a message to the point in time when it emits a message.

The early version of Erlang was developed by Ericsson Computer Science Laboratory in 1985. During that time, Ericsson couldn't find an appropriate language that has high performance in concurrency especially for telecommunication applications programming (for switching, trunking, etc), so they developed their own language. OTP stands for Open Telecom Platform, OTP was developed by Ericsson Telecom AB for programming next generation switches and many Ericsson products are based on OTP. OTP includes the entire Erlang development system together with a set of libraries written in Erlang and other languages. OTP was originally designed for writing telecoms application but has proved equally useful for a wide range of non-telecom that have concurrent, distributed, and also fault tolerant applications. In 1998 Ericsson released Erlang and the OTP libraries as open source. Now, Erlang/OTP has reached the R11 version.

III. WHY WE USE ERLANG?

The simple answer to the question above that is we need concurrency in the benchmark application. The application must be able to generate multiple users to do some stress tests to the web server.

But, there are some good features in Erlang, and even the other languages don't have these features. Some of these Erlang features are described below:

- 1) In Erlang processes are light weight.
- 2) Not only are Erlang processes light-weight, but also we can create many hundreds of thousands of such processes without noticeably degrading the performance of the system (unless of course they are all doing something at the same time)[5].
- 3) In Erlang, processes share no data and the only way in which they can exchange data is by explicit message passing. "dangling" pointers are very difficult to program in the presence of hardware failures - we took the easy way out, by disallowing all such data structures [4].
- 4) In Erlang, processes scheduling operation is done by its own virtual machine, so Erlang didn't inherit the underlying operating system processes scheduling. Real time. Erlang is intended for programming soft real-time systems where response times in the order of milliseconds are required.[4]
- 5) Continuous operation.
- 6) Automatic Memory management.
- 7) Distribution.

A. Concurrent and Distributed Erlang

Concurrent in Erlang involves processes creation and deletion. In order to create a new process in Erlang, we use BIF (Built In Function) *spawn/3* :

```
spawn(modulename,functionname,  
argumentlists)
```

or

```
pid_variabe=spawn(modulename,functionna  
me, argumentlists)
```

The illustration of process creation can be seen in figure 1.

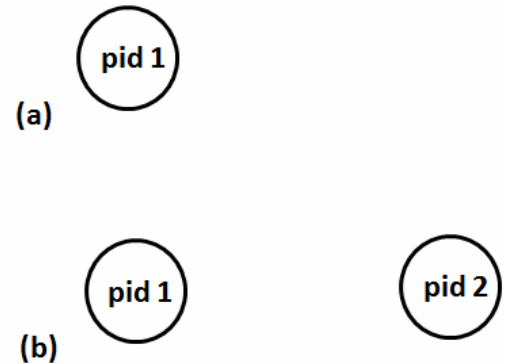


Figure 1. Process Creation Illustration

A process which no longer need by the system will be automatically shutdown/delete by the virtual machine (Erlang Runtime System/ERTS). Meanwhile, the message parsing mechanism can be done by these codes :

```
Pid ! Message  
.....  
Receive  
Message1 ->  
Actions1;  
Message2 ->  
Actions2  
.....  
After Time - >  
TimeOutActions  
end
```

The illustration for message parsing can be seen in figure 2.

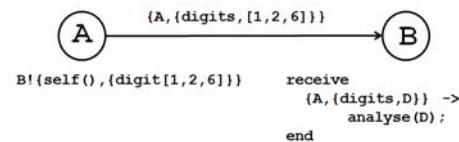


Figure 2. Message Parsing Illustration

Besides the example that has been shown above, in Erlang we can also use Behaviour in OTP Design Principles to create, delete and do message parsing between processes. A distributed Erlang system is a number of Erlang Runtime System (we called them nodes) that communicated each other by using message parsing with pid (process indentifier) through TCP/IP sockets transparently. A node must be given a

name before it can communicate each other. The name is either a long name or a short name like the examples below:

```
$erl -name dilbert[long name]
(dilbert@uab.ericsson.se)1>
@erl-sname dilbert[short
name](dilbert@uab)1>
```

A simple distribution in Erlang can be done by these codes:

```
...
Pid = spawn(Fun@Node)
...
alive(Node)
...
not_alive(Node)
```

B. Benchmark

The term benchmark can be described as:

- 1) A group of parameters in which products (software or hardware) can be measured the performance according to these parameters.
- 2) A computer program designed to measure the performance of software or hardware according to certain parameters.
- 3) A group of performance criteria that must be complied by software or hardware.

Web server benchmark means that a benchmark activity is made to the web server to measure its performance based on several parameters and using certain computer program to do this activity.

IV. DESIGNING

A. Application Concepts

Application is divided into two main parts, the controller section and the launcher section.

Controller is the core of the application. It has several duties, such as read the benchmark scenario file, configure the program based on the scenario, initialize the launcher section, gather the benchmark results from local and remote Erlang node where the launcher runs and write them in a log file (later the log file will be used to generate a report page for the sysadmin). Controller also has function as a timer which act as timing for user inters arrival to the server.

Launcher generates a number of user based on the scenario, initialize them and start the benchmark by sending requests to the web server. The clients also gather the benchmark result and send them to the controller. The illustration for Application Concepts can be seen in figure 3.

Several parameters that can be measured by this application are:

- 1) Number of Requests per second.
- 2) Simultaneous users that can be served by the server (per second).

- 3) Time that needs to connect for the client so it can be connected to server.
- 4) Number of user that can be served during the duration of benchmark.
- 5) Time that needs for a user so it can receive a full page of document/web page according to the request.
- 6) Time that needs to complete a session (a group of requests) as described in the scenario.
- 7) Network throughput.
- 8) HTTP Status (200, 404).

The application also has a report generator that written in PERL and using GNUPLOT to generate graphs based on the benchmark results from the log file.

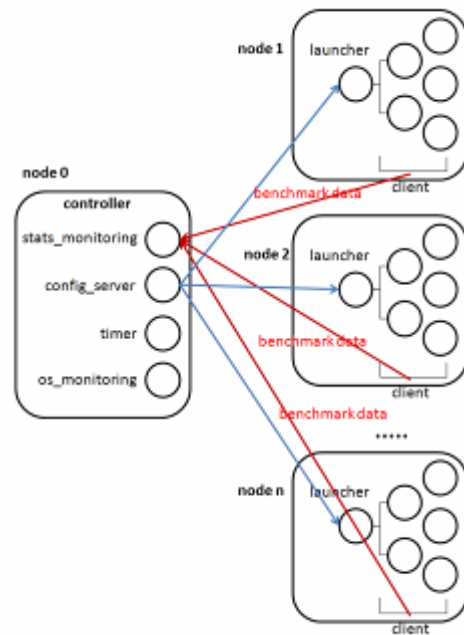


Figure 3. Illustration for Application Concepts

B. The Scenario Files

The benchmark scenario file is written using XML and consists of several sections:

- 1) The server section, where the user (sysadmin) describe the IP address of the server that he/she wants to benchmark.
- 2) The client section, in this section, user can write the IP addresses where the launcher section starts and generate a numbers of clients.
- 3) Inters arrival phase and benchmark duration section.
- 4) The simulated user agents (web browser) section.
- 5) The session and request section.
- 6) Each of the section describes above can be modified according to user necessity.

V. IMPLEMENTATION

A. Hardware and Software Specifications

Application is implemented by using a simple topology consists of two computers and a server across Local Area Network in Kapuk Valley, Margonda, Depok. The two computers using an Intel Celeron Processor (1.8 GHz and 2.28 GHz) and running Linux operating System (SuSE 10.0 and Slackware Linux 11.0), Open SSH v2, Erlang/OTP R11, PERL v5.8, and also BASH Shell. Both of them also connected to the TCP/IP network and within Kapuk-Valley domain (hostname mobile and posen-asik). The server using an Intel Pentium II 400 MHz. processor with 768 megabytes SDRAM memories and running Slackware Linux 11.0 with Apache 2.0.55 web server.

B. Testing and Implementation Process

In this implementation process, we're going to make a benchmark scenario with these parameters:

- 1) Total clients to generate: 600 clients.
- 2) Benchmark duration: 10 minutes.
- 3) Client inters arrival phase: 1 second.
- 4) 300 clients will be generating in host mobile, meanwhile the other 300 clients will be generate in host posen-asik.

Before we run the program, we must generate a pair of authentication key (public and private) for passwordless authentication using SHH in order to get the Erlang nodes to communicate each other.

```
[root@mobile~]#ssh-keygen -t rsa
Enter file in which to save the key
(/root/.ssh/id_rsa):
Enter passphrase (empty for no
passphrase):
Enter same passphrase again:
Your identification has been saved in
/root/.ssh/id_rsa.
Your public key has been saved in
/root/.ssh/id_rsa.pub.
The key fingerprint is:
ec:30:2c:c9:e0:0a:92:48:3c:e5:5a:f3:7c:69:
d8:92 root@mobile.myownlinux.org
[root@mobile~]#scp /root/.ssh/id_rsa.pub
root@posen-asik:/root/.ssh/
[root@posen-asik]#echo .ssh/id_rsa.pub >>
.ssh/authorized_keys}
```

After the process above, we can start the benchmark process by executing the shell script to initialize and start the application.

```
[root@mobile~]#./usr/local/bin/wii start
```

Several results from the important parameters to examine by the sysadmin are listed in the table I.

TABLE I
BENCHMARK RESULTS

Parameters	Result
Request per Second	3.4 requests per second
Connection per Second	1.8 connections per second
Page loaded per second	1.8 pages per second
Total user served	595 of 600 users

VI. CONCLUSIONS

After our studies we show that in Erlang:

- 1) Processes are light weight. Not only are Erlang processes light-weight, but also we can create many hundreds of thousands of such processes without noticeably degrading the performance of the system (unless of course they are all doing something at the same time),
- 2) Processes share no data and the only way in which they can exchange data is by explicit message passing. Erlang message never contain pointers to data and since there is no concept of shared data, each process must work on a copy of the data that it needs. All synchronization is performed by exchanging messages.
- 3) Processes scheduling operation is done by its own virtual machine.
- 4) Processes are real time. Erlang is intended for programming soft real-time systems where response times in the order of milliseconds are required.
- 5) Processes code has continuous operation. Erlang has primitives which allow code to be replaced in a running system and allow old and new versions of code to execute at the same time.
- 6) Processes operation use automatic memory management. Memory is allocated automatically when required, and deallocated when no longer used.
- 7) All interaction between processes is by asynchronous message passing. Distributed systems can easily be built.

By examine the results based on the important parameters; we hope that the syadmin and netadmin can make fine tuning to their server.

ACKNOWLEDGMENT

The authors would like to thank Gunadarma University's Research Council

REFERENCES

- [1] Anonymous, *Panduan Lengkap Pengembangan Jaringan Linux*, Penerbit Andi, Yogyakarta.
- [2] Anonymous, *Web Server*, [http://en.wikipedia.org/wiki/Web server](http://en.wikipedia.org/wiki/Web_server) (05 September 2010).

- [3] Anonymous, *An Erlang Course*, <http://www.erlang.org/download/course.pdf>, (05 September 2010).
- [4] Armstrong, Joe. et al., *Concurrent Programming in Erlang*, Prentice Hall, New Jersey.
- [5] Armstrong, Joe, *Concurrency Oriented Programming in Erlang*, Swedish Institute of Computer Science.
- [6] Cisco System, Cisco Networking Academy Program – CCNA Modules – 640-801, <http://cisco.netacad.net>, Washington.
- [7] Dudy Rudianto, *Desain dan Implementasi Sistem Operasi Linux*, Elex Media Komputindo, Jakarta, 2002.
- [8] Dudy Rudianto, *PERL Untuk Pemula*, Elex Media Komputindo, Jakarta.
- [9] Ericsson Telecommunications Systems Laboratories Team, *Erlang Reference Manual*, Ericsson Inc., Sweden.
- [10] Ericsson Telecommunications Systems Laboratories Team, *Getting Started With Erlang*, Ericsson Inc., Sweden.
- [11] Hedqvist, Pedka, *A Parallel and Multithreaded Erlang Implementation*, Uppsala University, Sweden.

AUTHORS PROFILE



A. Suhendra was born in Jakarta, 1969. He received the B.Sc. degree in Informatics Engineering from Gunadarma University, Depok, in 1992, and his MSc in Computer Science from AIT, Bangkok in 1994, and his Dr.-Ing. in Informatics from University of Kassel, Germany in 2002.



A.B. Mutiara was born in Jakarta, in 1967. He received the B.Sc. degree in Informatics Engineering from Gunadarma University, Depok Indonesia, in 1991, and the M.S. degree and Ph.D degree in computational material physics from Goettingen University, Germany, in 1997 and 2000.

Advanced Virus Monitoring and Analysis System

Fauzi Adi Rafrastara

Faculty of Information and Communication Technology
University of Technical Malaysia Melaka
Melaka, Malaysia
fauzi_adi@yahoo.co.id

Faizal M. A

Faculty of Information and Communication Technology
University of Technical Malaysia Melaka
Melaka, Malaysia
faizalabdollah@utem.edu.my

Abstract — This research proposed an architecture and a system which able to monitor the virus behavior and classify them as a traditional or polymorphic virus. Preliminary research was conducted to get the current virus behavior and to find the certain parameters which usually used by virus to attack the computer target. Finally, “test bed environment” is used to test our system by releasing the virus in a real environment, and try to capture their behavior, and followed by generating the conclusion that the tested or monitored virus is classified as a traditional or polymorphic virus.

Keywords—Computer virus, polymorphic virus, traditional virus, VMAS.

I. INTRODUCTION

Nowadays, we all live in the digital era, which most information moves from one place to another digitally. The information can be derived easily from everywhere and send it to whoever, only in minutes or even seconds. Unfortunately, wherever we are, including in this digital information era, threats always exist, perhaps in the different shapes. One of the popular threats which always peering us in this era, is Computer Virus.

The virus is a threat, because it can do bad things to whomever. It can make the computer becomes slow, broken, or even it can delete the data. The virus can run automatically and hide the process, so that users cannot see the processes and activities, which are done by virus. What can users see from the virus is what they have done.

II. BACKGROUND

There is a kind of software that can be used to detect the existence of Virus inside the computer, called Anti Virus (AV). AV is widely used to detect and combat the virus. They will report to the user when they found the virus inside the PC. Unfortunately, they cannot list and report all behaviors or activities of the virus [1]. This limitation of AV has been covered by the certain tools, which mostly do not have a capability in virus detection system, called Virus Monitoring and Analysis Tool (VMAS). VMAS is specially used to monitor and analyze as well as capture all activities performed by virus [1]. VMAS also can generate the details report regarding the virus's behavior. This kind of report is important for those who want to learn more about virus activities. Furthermore, people can eliminate the viruses from their PC

and recover the Operating System from viruses attack by reading the virus behavior analysis report [2]. There are several popular VMAS which mostly used to get the data of virus's behavior, such as CWSandbox, Capture, MBMAS, Joebox and ThreatExpert [2].

The aforementioned tools indeed are able to produce the behavior analysis report in details. Unfortunately, by using these tools, the type of malicious file, that have been tested, still cannot be recognized. Even though the analysis report can be derived, it is not easy to determine which virus file is classified as traditional or polymorphic only by reading this report [2][3].

However, either AV or VMAS cannot distinguish between traditional and polymorphic virus. They are only capable of detecting and reporting the virus behavior. Whereas, classifying the virus automatically, it will be a different task which has not been solved yet. So, in this research, a new architecture will be proposed as well as the system. This architecture and system are served to classify the virus automatically whether it is considered as a traditional or polymorphic virus.

III. DATA COLLECTION

Data Collection is needed to conduct the preliminary research in which all the required data will be collected manually. Further, these data will be compared to the generated data in testing phase. Here 20 viruses will be examined and analyzed one by one. This step is important to classify whether these viruses are categorized as a traditional or polymorphic virus. Based on this manual experiment, two viruses were detected as a polymorphic virus, since they always obfuscated their signatures whenever they propagate [4] [5], as listed in Table I. The signature that was identified in our research here is MD5 checksum [6][7]. This kind of checksum is popular to be used by current antivirus to detect the existence of viruses based on their signatures [8][9][10].

Further, these data will be used to validate the final data which generated by the proposed system. The proposed system can be considered to be successful if it can produce the same result and conclusion with the data from this preliminary research.

TABLE I. LIST OF THE ANALYZED VIRUS

No.	Virus Name	Detected by	Types
1.	W32.Blaster.E.Worm (Lovesan)	Symantec	Traditional
2.	W32.Downadup.B (Conficker)	Symantec	Traditional
3.	W32.Higuy@mm	Symantec	Traditional
4.	W32.HLLW.Benfgame.B (Fasong)	Symantec	Polymorphic
5.	W32.HLLW.Lovgate.J@mm	Symantec	Traditional
6.	W32.Imaut	Symantec	Traditional
7.	W32.Klez.E@mm	Symantec	Polymorphic
8.	W32.Kwbot.F.Worm	Symantec	Traditional
9.	W32.Mumu.B.Worm	Symantec	Traditional
10.	W32.Myto.b.AV@mm	Symantec	Traditional
11.	W32.SillyFDC (Brontok)	Symantec	Traditional
12.	W32.SillyFDC (Xema)	Symantec	Traditional
13.	W32.Sober.C@mm	Symantec	Traditional
14.	W32.Swen.A@mm	Symantec	Traditional
15.	W32.Valla.2048 (Xorala)	Symantec	Traditional
16.	W32.Virut.CF	Symantec	Traditional
17.	W32.Wullik@mm	Symantec	Traditional
18.	W32/Rontokbro.gen@MM	McAfee	Traditional
19.	W32/YahLover.worm.gen	McAfee	Traditional
20.	Worm:Win32/Orbina!rts	Symantec	Traditional

IV. THE PROPOSED ARCHITECTURE AND SYSTEM

Since the main objective of this research is to propose an architecture and system which is able to classify between traditional and polymorphic virus, so this research focuses on the host side attack only.

In this research, two tools have been developed to classify between polymorphic and traditional virus, which are *Virus Behavior Monitoring Tools* (VBMT) and *Virus Behavior Analysis Tool* (VBAT). These tools are included in one system, called *Advanced Virus Monitoring and Analysis System* (AVMAS).

VBMT is served to monitor the activity of virus. They will execute the virus and then captured all activities which are performed by virus, during monitoring time. Usually current VMASes take maximal 4 minutes along for the monitoring time [1][11][12]. The VBMT will be installed into two PCs. Later, the same virus will be executed and monitored inside these PCs, to know whether or not the virus performs different things, especially in term of offspring's signature.

On the other hand, VBAT is used to analyze the results that generated by each VBMT. This analysis process is important to come up with the conclusion that the tested virus is classified as a polymorphic or traditional virus.

The proposed architecture here actually can be implemented in two environments, which are real environment and virtual environment. Real environment means, by providing at least two PCs to test the virus and installing VBMT into these PCs. One more PC is needed to be installed with a VBAT. Fig. 1 illustrates the architecture of AVMAS in real environment.

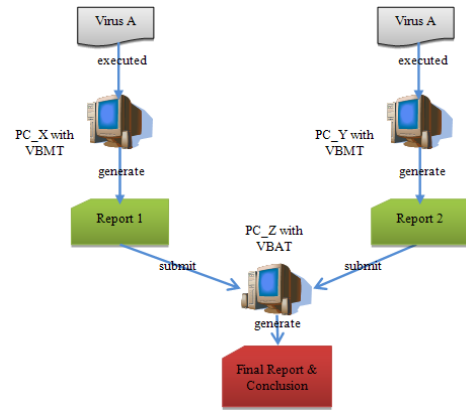


Figure 1. Architecture of AVMAS in real environment

The main concept of this architecture here is, a virus is tested in two PC with VBMT inside, by which VBMT will monitor and captured all activities which are performed by the tested virus. After monitoring time is finished, then each VBMT will generate a result that is reporting all activities captured, including new files generated and their checksums. Further, these two reports should be submitted to VBAT which installed inside the third PC. VBAT is tasked to analyze and compare between these two reports, and come up with the conclusion whether the tested virus is classified as a polymorphic or traditional virus. If VBAT found the fact that there is a difference between the first report with second report, especially in term of virus's activity or the signature of new files generated, so VBAT will conclude that the tested virus is classified as a polymorphic virus [1][4][5], otherwise it classified as a traditional virus [1][5].

This architecture actually can be simplified by using only one PC, but two virtual machines must be installed inside. The concept of the second architecture is almost similar to the first one. The difference is the location of VBMT which is installed inside these two virtual machines. Meanwhile, VBAT will be put inside the main or real PC. Fig. 2 shows the architecture of AVMAS in virtual environment.

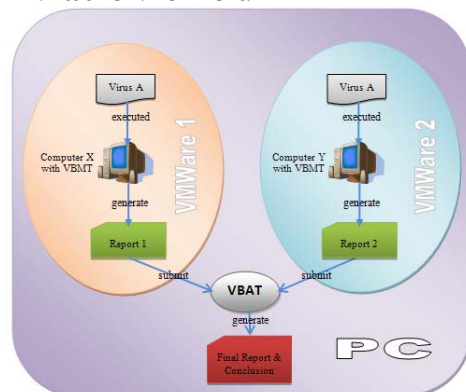


Figure 2. Architecture of AVMAS in virtual environment

V. TESTING AND RESULT

After completing the development phase, testing process should be done to make sure that the proposed system can be used to deal with the problems. Fig. 3 shows the flowchart to test the AVMAS. Firstly, a virus is put into two PCs with VBMT inside. After that, the virus is executed and monitoring process is started. Monitoring process will be performed in 5 minutes along, because according to [1][11][12], usually current VMAS take maximal 4 minutes to monitor virus activity. During this monitoring process, all virus behaviors, especially which relating to host side effect will be captured. When the timeout limit have been reached, each VBMT will generate the report consisting all behavior captured and the required data to classify the tested virus whether it is considered as a traditional or polymorphic virus.

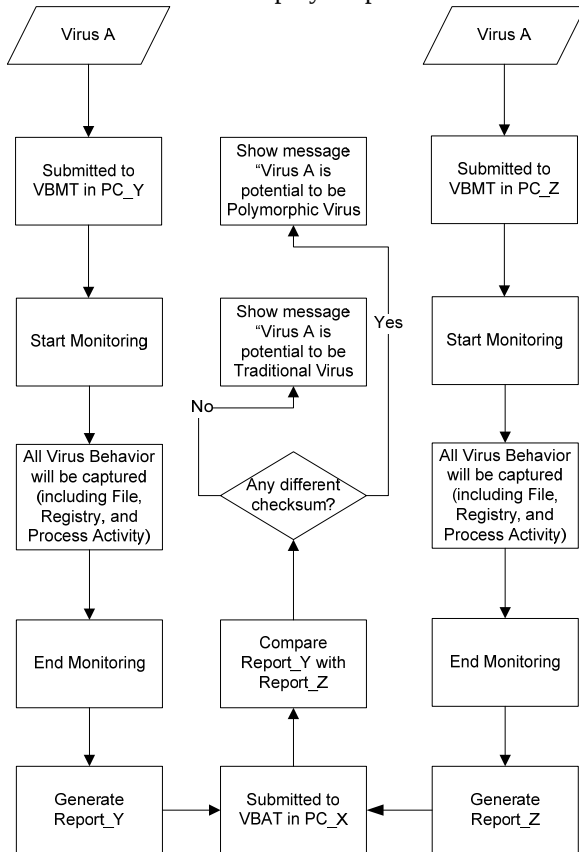


Figure 3. Flowchart to test the AVMAS

The next step is by submitting each report into VBAT which is installed in the third PC. This VBAT will compare the first report to the second report, especially in term of checksum generated. Once it finds the differences, so it means that, the tested virus can generate the different signature of offspring in the different PC. This conclusion addresses to the further conclusion that, this virus can be considered as a polymorphic virus.

On the other side, when the VBAT finds the same content between these two reports, including the generated checksums, so straight away VBAT will come up with the conclusion that this virus is classified as a traditional virus.

TABLE II. RESULT COMPARISON BETWEEN DATA FROM PRELIMINARY RESEARCH AND AVMAS TESTING

No.	Virus Name	Preliminary Research Result	AVMAS Testing Result
1.	W32.Blaster.E.Worm (Lovesan)	Traditional	Traditional
2.	W32.Downadup.B (Conficker)	Traditional	Traditional
3.	W32.Higuy@mm	Traditional	Traditional
4.	W32.HLLW.Benfgame.B (Fasong)	Polymorphic	Polymorphic
5.	W32.HLLW.Lovgate.J@mm	Traditional	Traditional
6.	W32.Imaut	Traditional	Traditional
7.	W32.Klez.E@mm	Polymorphic	Polymorphic
8.	W32.Kwbot.F.Worm	Traditional	Traditional
9.	W32.Mumu.B.Worm	Traditional	Traditional
10.	W32.MytoB.AV@mm	Traditional	Traditional
11.	W32.SillyFDC (Brontok)	Traditional	Traditional
12.	W32.SillyFDC (Xema)	Traditional	Traditional
13.	W32.Sober.C@mm	Traditional	Traditional
14.	W32.Swen.A@mm	Traditional	Traditional
15.	W32.Valla.2048 (Xorala)	Traditional	Traditional
16.	W32.Virut.CF	Traditional	Traditional
17.	W32.Wullik@mm	Traditional	Traditional
18.	W32.Rontokbro.gen@MM	Traditional	Traditional
19.	W32.YahLover.worm.gen	Traditional	Traditional
20.	Worm:Win32/Orbina!rts	Traditional	Traditional

Based on our test experiment, we found that this system can classify the tested virus correctly, with 100% similar to the data from preliminary research, as listed in Table II.

VI. CONCLUSION AND FUTURE WORK

In the monitoring process, this research focused on the host side attack, in which consist of three parameters that should be monitored, such as file, registry, and process activity. Whereas, to analyze the result for virus classification, there are several parameters used in this research, which are file activity, especially executable file creation, by comparing their checksums which produced in one PC to the checksum from another PC.

In the data collection phase, the viruses' behavior and activity especially which related to the host side have been captured, either manually or by using the third-party tools, such as: Joebox and ThreatsExpert. This data is used to match the result obtained from AVMAS. The result of this test and validation process show that, the system called AVMAS is able to monitor and classify the tested virus with same conclusion than one generated manually.

For the future work, this research can be improved to be a system, which is not only able to classify between traditional and polymorphic virus, but also to classify metamorphic virus as well. Next, this research also can be developed further to produce a system that is able to monitor and analyze the activity of a virus, then produce the virus removal tool automatically. It will be very beneficial to common users who want to clean their computers, which have been infected by the virus, since antivirus focuses on the prevention side so far, rather than cure action.

ACKNOWLEDGMENT

This research was supported by University of Technical Malaysia Melaka (UTeM) and Fundamental Research Grant Scheme (FRGS).

REFERENCES

- [1] Rafrastara, FA (2010). "Constructing Polymorphic Virus Analysis Tool Using Behavior Detection Approach," Master's Thesis. University of Technical Malaysia Melaka.
- [2] FuYong, Z, DeYu, Q, & JingLin, H (2005). "MBMAS: A System for Malware Behavior Monitor and Analysis." *CNMT '09* (pp. 1-4). Wuhan: IEEE.
- [3] Bayer, U (2005). "TTAnalyze: A Tool for Analyzing Malware." Master's Thesis. Technical University of Viena.
- [4] Abdulla, SM & Zakaria, O (2009). "Devising A biological model to detect polymorphic computer viruses." *ICCTD '09*, vol. 1, (pp. 300-304). Kota Kinabalu: IEEE.
- [5] Szor, Peter (2005). "The Art of Computer Virus Research and Defense." Maryland: Addison Wesley Profesional.
- [6] Stallings, William (2003). "Cryptography and Network Security: Principles and Practices, Third Edition." New Jersey, USA: Pearson Education, Inc.
- [7] Symantec. *MD5 Hash*. [Online] Retrieved on October 2010 from http://us.norton.com/security_response/glossary/define.jsp?letter=m&word=md5-hash.
- [8] ClamAV (2010). *Creating Signatures for ClamAV*. [Online] Retrived on October 2010 from <http://www.clamav.net/doc/latest/signatures.pdf>.
- [9] Zhou, X, Xu, B, Qi, Y, & Li, J (2008). "MRSI: A Fast Pattern Matching Algorithm for Anti-Virus Applications." *Seventh International Conference on Networking*. IEEE.
- [10] Zidouemba A (2009). *Writing ClamAV Signatures*. [Online] Retrieved on October 2010 from <http://www.clamav.net/doc/webinars/Webinar-Alain-2009-03-04.pdf>.
- [11] Bayer, U, Kirda, E, & Kruegel, C (2010). "Improving the Efficiency of Dynamic Malware Analysis." *SAC '10* (pp. 1871-1878). Sierre: ACM.
- [12] Rafrastara, FA & Abdollah, MF (2010). "Penetrating the Virus Monitoring and Analysis System Using Delayed Trigger Technique." *ICNIC'10*. IEEE.

AUTHORS PROFILE

Fauzi Adi Rafrastara. He was born in Semarang, Indonesia, 30 April 1988. He obtained the bachelor's degree from Dian Nuswantoro University, Indonesia, in 2009. He is currently a master student at University of Technical Malaysia Melaka. His research area include software engineering and information security.

Faizal M. A. He is currently a senior lecturer in University Technical Malaysia Melaka. He obtained the Ph.D degree in 2009 from University Technical Malaysia Melaka focusing on Intrusion Detection System. He research are IDS, Forensic, and Network Security.

A New Approach for Clustering Categorical Attributes

Parul Agarwal¹, M. Afshar Alam²

Department Of Computer Science, Jamia
Hamdard(Hamdard University)
Jamia Hamdard(Hamdard University)
New Delhi =110062 ,India

parul.pragna4@gmail.com, aalam@jamiahamdard.ac.in

Ranjit Biswas³

Manav Rachna International University
Manav Rachna International University
Green Fields Colony
Faridabad, Haryana 121001

ranjitbiswas@yahoo.com

Abstract— Clustering is a process of grouping similar objects together and placing the object in a cluster which is most similar to it. In this paper we provide a new measure for calculation of similarity between 2 clusters for categorical attributes and the approach used is agglomerative hierarchical clustering .

Keywords- Agglomerative hierarchical clustering, Categorical Attributes, Number of Matches.

I. INTRODUCTION

Data Mining is a process of extracting useful information. Clustering is the problem being solved in data mining. Clustering discovers interesting patterns in the underlying data. It groups similar objects together in a cluster (or clusters) and dissimilar objects in other cluster (or clusters). This grouping is based on the approach used for the algorithm and the similarity measure which identifies the similarity between an object and a cluster. The approach is based upon the clustering method chosen for clustering. The clustering methods are broadly divided into hierarchical and partitional. Hierarchical clustering performs partitioning sequentially. It works on bottom-up and top-down. The bottom up approach known as agglomerative starts with each object in a separate cluster and continues combining 2 objects based on the similarity measure until they are combined in one big cluster which consists of all objects. Whereas the top-down approach also known as divisive treats all objects in one big cluster and the large cluster is divided into small clusters until each cluster consists of just a single object. The general approach of hierarchical clustering is in using an appropriate metric which measures distance between 2 tuples and a linkage criteria which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. The linkage criteria could be of 3 types [28] single linkage, average linkage and complete linkage.

In single linkage (also known as nearest neighbour), the distance between 2 clusters is computed as:

$$D(C_i, C_j) = \min \{D(a, b) : \text{where } a \in C_i, b \in C_j\}.$$

Thus distance between clusters is defined as the distance between the closest pair of objects, where only one object from each cluster is considered.

i.e. the distance between two clusters is given by the value of the shortest link between the clusters. In average Linkage method (or farthest neighbour), Distance between Clusters defined as the distance between the most distant pair of objects, one from each cluster is considered.

In the complete linkage method, $D(C_i, C_j)$ is computed as

$$D(C_i, C_j) = \text{Max} \{ d(a, b) : a \in C_i, b \in C_j \}$$

the distance between two clusters is given by the value of the longest link between the clusters.

Whereas, in average linkage

$D(C_i, C_j) = \{ d(a, b) / (l_1 * l_2) : a \in C_i, b \in C_j \}$. And l_1 is the cardinality of cluster C_i , and l_2 is cardinality of Cluster C_j .

And $d(a, b)$ is the distance defined. }

The partitional clustering on the other hand breaks the data into disjoint clusters. In Section II we shall discuss the related work. In Section III, we shall talk about our algorithm followed by section IV containing the experimental results followed by Section V which contains the conclusion and Section VI will discuss the future work.

II. RELATED WORK

The hierarchical clustering forms its basis with older algorithms Lance-Williams formula (based on the Williams dissimilarity update formula which calculates dissimilarities between a cluster formed and the existing points, which are based on the dissimilarities found prior to the new cluster), conceptual clustering, SLINK[1], COBWEB[2] as well as newer algorithms like CURE[3] and CHAMELEON[4]. The SLINK algorithm performs single-link (nearest-neighbour) clustering on arbitrary dissimilarity coefficients and constructs a representation of the dendrogram which can be

converted into a tree representation. COBWEB constructs a dendrogram representation known as a classification tree that characterizes each cluster with a probabilistic distribution. CURE(Clustering using Representatives) an algorithm that handles large databases and employs a combination of random sampling and partitioning. A random sample is drawn from the data set and then partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters. CURE has the advantage of effectively handling outliers. CHAMELEON combines graph partitioning and dynamic modeling into agglomerative hierarchical clustering and can perform clustering on all types of data. The interconnectivity between two clusters should be high as compared to intra connectivity between objects within a given cluster..

Whereas, in the partitioning method, a partitioning algorithm arranges all the objects into various groups or partitions, where the total number of partitions(k) is less than the number of objects(n).i.e. a database of n objects can be arranged into k partitions, where $k < n$. Each of the partition thus obtained by applying some similarity function is a cluster. The partitioning methods are subdivided as probabilistic clustering[5] (EM ,AUTOCLASS), algorithms that use the k -medoids method (like PAM[6], CLARA[6],CLARANS[7]), and k -means methods (differ on parameters like initialization, optimization and extensions).EM (expectation – maximization algorithm) calculates the maximum likelihood estimate by using the marginal likelihood of the observed data for a given statistical model which depends on unobserved latent data or missing values .But this algorithm depends on the order of input. AUTOCLASS algorithm works for both continuous and categorical data. AUTOCLASS, is a powerful unsupervised Bayesian classification system which mainly has application in biological sciences and is able to handle the missing values. PAM (partitioning around medoids) builds k representative objects, called medoids randomly from given dataset consisting of n objects . A medoid is an object of a given cluster such that its average dissimilarity to all the objects in the cluster is the least. Then each object in the dataset is assigned to the nearest medoid. The purpose of the algorithm is to minimize the objective function which is the sum of the dissimilarities of all the objects to their nearest medoid.

CLARA (Clustering Large Applications) deals with large data sets.it combines sampling and PAM algorithm to to generate an optimal set of medoids for the sample. It also tries to find k representative objects that are centrally located in the cluster.It considers data subsets of fixed size, so that the overall computation time and storage requirements become linear in the total number of objects. CLARANS (Clustering Large Applications based on RANdomized Search) views the process of finding k medoids as searching in a graph [12]. CLARANS performs serial randomized search instead of exhaustively searching the data.It identifies spatial structures present in the data.

Partitioning algorithms are also density based i.e. try to discover dense connected components of data, which are flexible in terms of their shape. Several algorithms like DBSCAN[8], OPTICS have been proposed.. The DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)algorithm identifies clusters on the basis of the density of the points.

Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or outliers. Its main features include ability to handle large datasets with noise,identifying clusters with different sizes and shapes.OPTICS (Ordering Points To Identify the Clustering Structure) though similar to DBSCAN in being density based and working over spatial data but differs by considering the problem posed by DBSCAN problem of detecting meaningful clusters in data of varying density.

Another category is grid based methods like BANG[9] in addition to evolutionary methods such as Simulated Annealing(a probabilistic method of calculating the global minimum over a cost function having many local minimas),Genetic Algorithms[10].Several scalability algorithms e.g. BIRCH[11],DIGNET[12] have been suggested in the recent past to address the issues associated with large databases . BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an incremental and agglomerative hierarchical clustering algorithm for databases which are large enough to not fit the main memory. This algorithm performs only single scan of the database and effectively deals with data containing noise.

Another category of algorithms deals with high dimensional data and works on Subspace Clustering,Projection Techniques,Co-Clustering Techniques. Subspace clustering finds clusters in various subspaces within a dataset. High dimensional data may consist of thousands of dimensions and thus may pose difficulty in their enumeration due to their multiple values that they may take and visualization owing to the fact that many of the dimensions may often be irrelevant.. The problem with subspace clustering is ,that with d dimensions there exist 2^d subspaces.Projected clustering[14] assigns each point to a unique cluster, but clusters may exist in different subspaces. Co-Clustering or Bi-Clustering[15] is simultaneous clustering of rows and columns of a matrix i.e. of tuples and attributes.

The techniques of grouping the objects are different for numerical and categorical data owing to their separate nature. The real world databases contain both numerical and categorical data.Thus, we need separate similarity measures for both types. The numerical data is generally grouped on the basis of the inherent geometric properties like distances(most common being Euclidean, Manhattan etc) between them. Whereas for categorical data the attribute values that they take is small in number and secondly, it is difficult to measure their similarity on the basis of the distance as we can for real numbers. There exist two approaches for handling mixed type of attributes. Firstly, group all the same type of variables in a particular cluster and perform separate dissimilarity computing method for each variable type cluster. Second approach is to group all the variables of different types into a single cluster using dissimilarity matrix and make a set of common scale variables. Then using the dissimilarity formula for such cases, we perform the clustering.

There exist several clustering algorithms for numerical datasets. The most common being K-means, BIRCH, CURE, CHAMELEON. The k-means algorithm takes as input the number of clusters desired. Then from the given database, it randomly selects k tuples as centres and then assigning the objects in the database to belong to these clusters on the basis of the distance. It then recomputes the k centres and continues the process till the centres don't move. K-means was further proposed as fuzzy k-means and also for categorical attributes. The original work has been explored by several authors for extension and several algorithms for the same have been proposed in the recent past. In [16], Ralambondrainy proposed an extended version of the k-means algorithm which converts categorical attributes into binary ones. In this paper the author represents every attributes in the form of binary values which results in increased time and space incase the number of categorical attributes is large.

A few algorithms have been proposed in the last few years which cluster categorical data. A few of them listed in [17-19]. Recently work has been done to define a good distance (dissimilarity) measure between categorical data objects [20-22, 25]. For mixed data types a few algorithms [23-24] have been written. In [22] the author presents k-modes algorithm, an extension of the K-means algorithm in which the number of mismatches between categorical attributes is considered as the measure for performing clustering. In k-prototypes, the distance measure for numerical data is weighted sum of Euclidean distances and for categorical data, a measure has been proposed in the paper. K-Representative is a frequency based algorithm which considers the frequency of attribute in that cluster and dividing it by the length of the cluster.

III. The Proposed Algorithm (namely MHC) (Matches based Hierarchical-Clustering) where M stands for the number of matches.

This algorithm works for categorical datasets and constructs clusters hierarchically.

Consider a database D . If D is the database with domain $D_1, \dots, D_m$ defined by attributes $A_1, \dots, A_m$, then each tuple X in D is represented as

$$X = (x_1, x_2, \dots, x_m) \in (D_1 \times D_2 \times \dots \times D_m). \quad (1)$$

Let, there be n objects in the database, then

$$D = (X_1, X_2, \dots, X_n). \text{ Where object } X_i \text{ is represented as } X_i = (x_{i1}, x_{i2}, \dots, x_{im}) \quad (2)$$

Where m is the total number of attributes. Then, we define similarity between any two clusters as

$$\text{Sim}(C_i, C_j) = \text{matches}(C_i, C_j) / (n * (l_i * l_j)) \quad (3)$$

Where C_i, C_j denote the clusters for which similarity is being calculated.

$\text{matches}(C_i, C_j)$: denote the number of matches between 2 tuples over corresponding attributes.

n : total number of attributes in database

l_i : length of the C_i cluster

l_j : length of the C_j cluster

A. The Algorithm:

Input: Number of Clusters (k), Data to be Clustered (D)

O/p: k number of clusters created.

Step 1. Begin with n Clusters, each having just one tuple.

Step 2. Repeat step 3 for $n-k$ times.

Step 3. Find the most similar cluster C_i and C_j using the similarity measure $\text{Sim}(C_i, C_j)$ by "(3)" and merge them into a single cluster.

B. Implementation Details:

1. This algorithm has been implemented in Matlab [26] and the main advantage is that we do not have to reconstruct the similarity matrix once this task is done.

2. It is simple to implement.

3. Given n tuples construct $n * n$ similarity matrix with all $i=j$ value initially set to 8000 (a special value) and the rest with a value 0.

4. During 1st iteration, calculate the similarity of each cluster with every other cluster. for all i, j s.t. $i \neq j$. Compute the similarity between 2 tuples (clusters) of database by identifying the number of matches over attributes and then using equation (3) to calculate the value for this step and accordingly update the matrix.

5. Since only the upper triangular matrix will be used, identify the highest value from matrix and merge the corresponding i and j . the changes in the matrix include:

a) set $(j, j) = -9000$ to identify that this cluster has been merged with some other cluster.

b) set $(i, j) = 8000$ which denotes that for corresponding row i , all j 's with 8000 as value have been merged with i .

c) During next iteration, do not consider the similarity between those clusters which have been merged. for example if database D contains 4 (n) tuples with 5 (m) attributes, and 1, 2 have been merged then following similarities have to be calculated.

$$\text{sim}(1, 3) = \text{sim}(1, 3) + \text{sim}(2, 3) \text{ where } l_i = 2, l_j = 1$$

$$\text{sim}(3, 4) \text{ where } l_i = 1, l_j = 1$$

$$\text{sim}(1, 4) = \text{sim}(1, 4) + \text{sim}(2, 4) \text{ where } l_i = 2, l_j = 1$$

IV. EXPERIMENTAL RESULTS

We have implemented this algorithm with small size synthetic database and the results have been good. But as the size increases, the algorithm has the drawback of producing mixed clusters. Thus, we consider a real life dataset which is small in size for experiments.

Real life dataset:

This dataset has been taken from UCI machine learning repository[27]The dataset is the soyabean small dataset.

A small subset of the original soybean (large)database.The soyabean large has 307 instances and 35 attributes alongwith some missing values.The data has been classified into 19 classes.On the other hand,the soyabean small dataset with no missing values consisting of 47 tuples with 35 attributes . The dataset has been classified into 4 classes.Both the datasets are being used for soyabean disease diagnosis.A few of the attributes are germination(in %),area damaged, plant growth(norm,abnorm),leaves(norm,abnorm),etc

Table 1

Classes	Expected No. Of Clusters	Resultant No. of Clusters
1	10	10
2	10	10
3	10	10
4	17	17

A. Validation Methods:

1.Precision(P): Precision in simplest terms can be formulated as number of objects identified correctly which belong to the class divided by the number of objects identified in that class.

2.Recall (R): Recall can be formulated as the number of objects correctly identified in that class divided by the total number of objects this class correctly has.

3. F measure (say denoted by F):it is the harmonic mean of precision and recall.

$$\text{i.e. F-Measure} = (2 * P * R) / (P + R). \quad (4)$$

The following Tables contain the values of the three validation measures discussed above for the algorithms ROCK,K-modes with our algorithm.

We assign the four classes obtained in results as c1,c2,c3,c4 and the actual classes as C1,C2,C3,C4.

Table 2 (MHC)

	C1	C2	C3	C4	P	R	F
c1	10	0	0	0	1	1	1
c2	0	10	0	0	1	1	1
c3	0	0	10	0	1	1	1
c4	0	0	0	17	1	1	1

Table 3 ROCK

	C1	C2	C3	C4	P	R	F
c1	7	0	0	8	0.47	0.70	0.56
c2	1	7	0	0	0.87	0.70	0.78
c3	1	3	4	0	0.50	0.40	0.44
c4	1	0	6	9	0.56	0.52	0.55

Table 4 k-Modes

	C1	C2	C3	C4	P	R	F
c1	2	0	0	7	0.22	0.20	0.21
c2	0	8	1	0	0.89	0.80	0.84
c3	6	0	7	1	0.50	0.70	0.58
c4	2	2	2	9	0.60	0.52	0.56

Using Table 4, we shall show how to calculate Precision,Recall and F-Measure for a particular class say c1 for k –modes.

In class c1, there are 2 tuples that actually belong to c1 and 7 tuples that belong to class c4.

So, Precision(P) = $2/(2+7)=0.22$

Also, there should be total of 10 tuples that should belong to this class against 2 which have been obtained by k modes algorithm

So, Recall(R) = $2/10 = 0.20$

F-Measure calculated using eqn (4) for class c1
= $(2*0.22*0.20)/(0.22+0.20) = 0.21$

Thus experimental results clearly indicate that M-Cluster has generated accurate achieving 100 % accuracy in contrast to other algorithms.

V. CONCLUSION

This algorithm produces good results for small databases. The advantages are that it is extremely simple to implement, memory requirement is low and accuracy rate is high as compared to other algorithms.

VI FUTURE WORK

We would like to analyse the results for large databases as well.

REFERENCES

- [1] SLINK: An optimally efficient algorithm for the single link cluster method. The Computer Journal, 16(1):30-34.
- [2] Fisher, Douglas H. (1987). "Knowledge acquisition via incremental conceptual clustering". Machine Learning 2: 139-172
- [3] Sudipto Guha,Rajeev rastogi,Kyuseok Shim:"CURE": A Clustering Algorithm for large Databases. Proc. of the ACM_sigmod Int'l Conf. .
- [4] Clustering Algorithm using dynamic Modelling.IEEE Computer ,/*Vol. 32.No. 8,68-75,1999.
- [5] T. Mitchell, Machine Learning, McGraw Hill, 1997.
- [6] Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data—An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. NewYork: JohnWiley & Sons, Inc.
- [7] Ng, R. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining.In Proceedings of the 20th international conference on very large data bases, Santiago,Chile, pages 144-155. Los Altos, CA: Morgan Kaufmann.
- [8] Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). Adensity-based algorithm for discovering clusters in large spatial databases with noise. In Second international conference on knowledge discovery and data mining, pages 226-231. Portland, OR: AAAI Press.
- [9] Schikuta, E. and Erhart, M. (1997). The BANG-clustering system: Grid-based data analysis.In Liu, X., Cohen, P., and Berthold, M., editors, Lecture Notes in Computer Science, volume 1280, pages 513-524. Berlin, Heidelberg: Springer-Verlag.
- [10] [Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning.Reading, MA: Addison-Wesley.
- [11] [T.Zhang,R.Ramakishnan,M.livny:" BIRCH:An Efficient data Clustering method for very large databases.Proc. of the ACM_sigmod International Conference on Management of data,1996,pp.103-114
- [12] R. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In VLDB-94, 1994.
- [13] Kriegel, Hans-Peter; Kröger, Peer; Renz, Matthias; Wurst, Sebastian (2005), "A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data", Proceedings of the Fifth IEEE International Conference on Data Mining (Washington, DC: IEEE Computer Society): 205-25
- [14] [Aggarwal, Charu C.; Wolf, Joel L.; Yu, Philip S.; Procopiuc, Cecilia; Park, Jong Soo (1999), "Fast algorithms for projected clustering", ACM SIGMOD Record (New York, NY: ACM) 28 (2): 61-72,
- [15] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," IEEE/ACM Trans. Computat. Biol. Bioinformatics, vol. 1, no. 1, pp. 24-45, Jan. 2004.
- [16] H.Ralambondrainy,"Aconceptual versionof the kmeans al gorithm"
- [17] .ArnRecogn. Lett., Vol. 15, No. 11, pp. 1147-1157, 1995.
- [18] Sudipto Guha,Rajeev rastogi,Kyuseok Shim:"ROCK":A robust clustering algorithm for categorical attributes".In Proc. 1999 International Conference of . data Engineering,pp.512-521,Sydney,Australia,Mar.1999.
- [19] [Yi Zhang,Ada Wai-chee Fu,Chun Hing Cai,peng-Ann Heng:"Clustering Categorical Data.In Proc. 2000 IEEE Int. Conf. Data Engineering,San Diego,USA,March 2000.
- [20] David Gibson,Jon Klieberg, Prabhakar Raghavan,"Clustering Categorical Data :An Approach based on Dynamic Syatems.Proc. 1998. Int. Conf. On Very Large Databases,pp. 311-323,New York,August 1998.
- [21] U.R. Palmer,C.Faloutsos.Electricity based External Similarity of Categorical Attributes.In Proc. Of the 7th Pacific Asia Conference on Advances in Knowledge Discovery and Data Mining PAKDD'03,pp. 486-500,2003.
- [22]] Z. Huang, "Extensions to the k-means algorithm for clustering large datasets with categorical values", Data Mining Knowl. Discov., Vol. 2, No. 2
- [23] C.Li,G. Biswas. Unsupervised learning with mixed Numeric and nominal data.IEEE Transactions on Knowledge and Data Engineering,2002,14(4):673-690.
- [24] S-G. Lee,D-K. Yun.Clustering categorical and Numerical data: A New procedure using Multi Dimensional scaling.International Journal of Information Technology and Decision Making, 2003,2(1) :135-160.
- [25] Ohn Mar San, Van-Nam Huynh, Yoshiteru Nakamori, "An Alternative Extension of The K-Means algorithm For Clustering Categorical Data", J. Appl. Math. Comput. Sci, Vol. 14, No. 2, 2004, 241-247.
- [26] MATLAB. User's Guide. The Math- Works, Inc., Natick, MA 01760, 1994-2001.
<http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml>
- [27] .P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1992. www.ics.uci.edu/~mllearn/MLRepository.html
- [28] www.resample.com/xlminer/help/HClst/HClst_intro.htm

Position based Routing Scheme using Concentric Circular Quadrant Routing Protocol in mobile ad hoc network

Upendra Verma

M.E. (Research Scholar)

Shri Vaishnav Institute Of Science and
Technology Indore(Madhya Pradesh),INDIA
upendra4567@gmail.com

Mr. Vijay Prakash

Asst. Prof.

Shri Vaishnav Institute Of Science and
Technology Indore(Madhya Pradesh),INDIA
vijayprakash15@gmail.com

Abstract— Mobile ad hoc network is a self-organizing and self-configuring network in which, mobile host moved freely, due to this disconnection is often occurred between mobile hosts. In mobile ad hoc network location of mobile nodes are frequently changed. Location management is a crucial issue in manet due to dynamic topologies. In this paper we propose a position based routing scheme called “Concentric Circular Quadrant Routing Scheme (CCQRS)”. In this paper we use single location server region for location update and location query. This strategy reduces the cost associated with location update and location query.

Keywords- mobile adhoc network; routing protocol; Concentric Circular location management;CCQRS.

I. INTRODUCTION

Mobile ad hoc network is self configuring network using network of mobile nodes connected by wireless link. In manet, mobile nodes are free to join or leave the network and they move randomly. Due to this, the network topology is frequently changed, that means dynamic network topology is used in manet. Mobile ad hoc network is highly suited for use in situations where fixed infrastructure is not available, not trusted, too expensive or unreliable. In manet, there is no need for planning of base station installation or wiring. In manet, users accomplishing their task, accessing information and communicating anytime, anywhere and from any device or node.

There are various application of manet. The original motivation of manet for military application. In battlefield, military can not rely on access to fixed infrastructure. Manet is also used for emergency services such as search and rescue operation, disaster recovery, replacement of fixed infrastructure in case of environmental disaster, policing and fire fighting, supporting doctors and nurses in hospital. Manet is also used for conference and meeting routs, office wireless networking, network at construction sites, inter vehicle network. Manet is also used for education such as virtual classrooms, ad hoc communicating during meetings and lectures. Manet is used for entertainment like multi-user

games. Manet is also used for location specific services, time dependent services.

In this scheme, we assumed that the whole network area is a circular. Here this circular area contains another circle, which is called location server region. We use only one location server region in which nodes (location servers) are relative fixed.

The rest of paper is organized as follows. In section 2 , Classification of Routing Protocol are presented. Section 3 gives description of Proposed scheme. Section 4 gives conclusion.

II. CLASSIFICATION OF ROUTING PROTOCOL

Routing is the act of moving information from source to destination in internetwork. During this process, at least one intermediate node within the internetwork is encountered.

Problem with routing in mobile ad hoc network:

– **Asymmetric links:** Most of the wired networks rely on the symmetric links which are always fixed. But this is not a case with ad-hoc networks as the nodes are mobile and constantly changing their position within network. For example consider a MANET(Mobile Ad-hoc Network) where node B sends a signal to node A but this does not tell anything about the quality of the connection in the reverse direction.

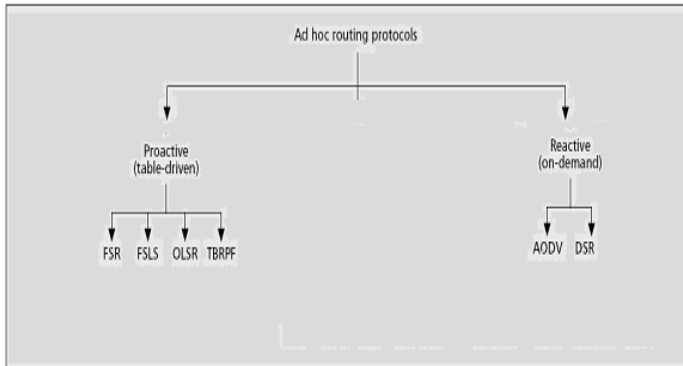
– **Routing Overhead:** In wireless adhoc networks, nodes often change their location within network. So, some stale routes are generated in the routing table which leads to unnecessary routing overhead.

– **Interference:** This is the major problem with mobile ad-hoc networks as links come and go depending on the transmission characteristics, one transmission might interfere with another one and node might overhear transmissions of other nodes and can corrupt the total transmission.

– **Dynamic Topology:** This is also the major problem with ad-hoc routing since the topology is not constant. The mobile node might move or medium characteristics might change. In

ad-hoc networks, routing tables must somehow reflect these changes in topology and routing algorithms have to be adapted. For example in a fixed network routing table updating takes place for every 30sec. This updating frequency might be very low for ad-hoc networks.

There are two types of routing protocol- Table driven routing protocol (Proactive) and on demand routing protocol (Reactive) as shown in table:



In proactive protocol each and every node in the network maintains routing information of every other node in the network. Routing information is generally kept in the routing tables and it is periodically updated as network topology changes. Proactive protocol is suitable for small network not for larger network because protocol need to maintains node entries for each and every node in the routing table. This causes more overhead in the routing table leading to consumption of more bandwidth.

Reactive Protocol is a protocol, they don't maintain routing information. If a node want to send message to another node then this protocol searches for route in on demand manner and establish connection in order to transmit and receive the message. The route discovery occurs by flooding the route request packets throughout the network.

here, we will use DSDV (Distance sequenced distance vector) Protocol. DSDV is a proactive protocol, which is based on bellman ford algorithm. It was developed by c peakins and P. bhagwat in 1994. DSDV solve the routing loop problem. For routing loop problem, each entry in the routing table contains a sequence number.

III. DESCRIPTION OF PROPOSED SCHEME

A position based routing scheme called "Concentric Circular Quadrant Routing Scheme (CCQRS)" is used for location update and location query. In this scheme I have assumed that whole network area as circular. In this scheme a small circle region called location server region denoted by **R** is in the other circular area. This circular area divides

into four quadrant. Each four quadrants uses the small circle region i.e. location server region for location update and location query. In this scheme only one location server region is used. Location server region contains multiple location servers.

This scheme uses "one for all location service", i.e. one location sever region (containing nodes acting as location server) is used for location update and location query.

The architecture of Concentric Circular Quadrant as shown in fig.

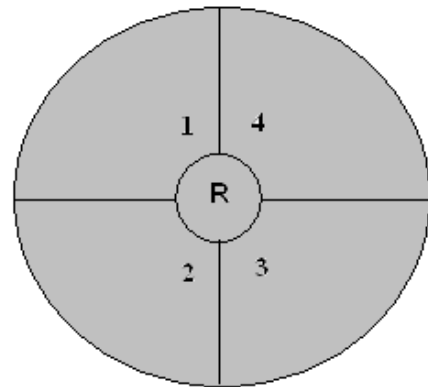


Figure 1 Concentric Circular Quadrant Architecture

Here, only one location server region say **R**. It contains many nodes acting as location servers. This location server region will stores complete location information of a node. Complete location information consists of nodeid, quadrant number, x-coordinate, y-coordinate.

A. Location Server Update:

Let 'p' be a node in the network. There are two movement of p:

- movement within the region (Quadrant)
- movement between the region(Quadrant).

i) Movement within the region(Quadrant):-

Whenever node 'p' moves from one location to other location within the same quadrant. In this situation, after movement node p informs to location server. Location server updates the existing location information related to node p as shown fig.

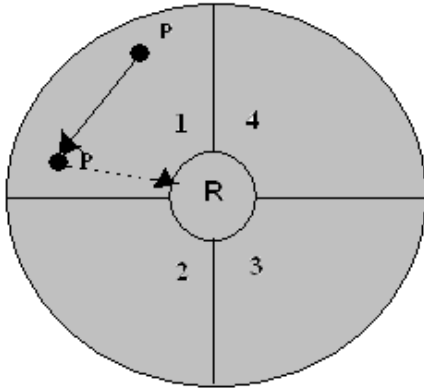


Figure 2: movement within the quadrant

ii) Movement between the region(Quadrant):-

Whenever node 'P' moves from one location to other location between the quadrants (i.e. different quadrant). In this situation, after movement node P informs to same location server means that in this situation same location server region is used. Location server updates the existing location information related to node P as shown in fig.

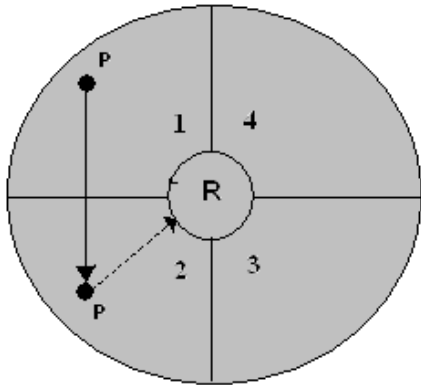


Figure 3: movement between the quadrant

B. Location query:

There are two cases for location query:

- i) Location query within the quadrant
- ii) Location query between the quadrants

i) Location query within the quadrant:

Node P sends a message to node R. in this situation node P sends the message to one of the location servers, which is in location server region. The location server contains the complete information for node R. Location server sends message directly to node R as shown in fig.

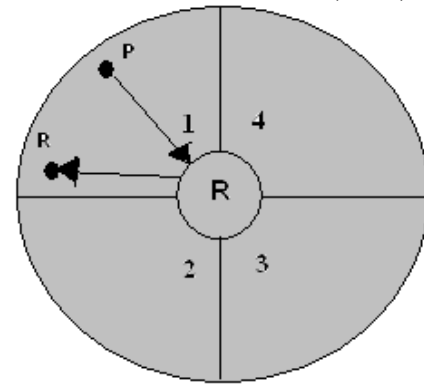


Figure 4: Location query within the quadrant

ii) Location query between the quadrants:

Node P sends a message to node R. here, node P is in a quadrant and node R is in other quadrant. In this situation node P sends the message to one of the location servers, which is in location server region. The location server contains the complete information for node R. Location server sends message directly to node R as shown in fig. Same procedure for both queries because of only one location server region.

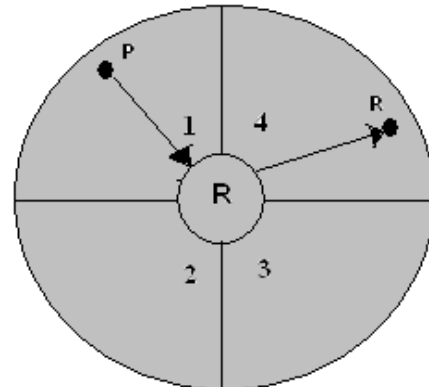


Figure 5: Location query between the quadrants

IV. PROPOSED ALGORITHMS

- Algorithm for movement of nodes within the quadrant

Let us take: Q0,Q1,Q2,Q3 [for quadrants]
P [for node]

1. *begin*: node p move from one location to another location [Q0toQ0 or Q1toQ1 etc.] within the quadrant
2. *move*: moverment of nodes within the quadrant. After movement the information of node (radius r) is changed in location server of location server region.

3. *protocol*: for communication any one of the protocols is used (e.g. DSDV, AODV, DSR, LAR etc)
 4. *store*: New information is stored into the location server
 5. *repeat*: [step 1 to step 2] for new moving nodes
- Algorithm for movement of nodes across the quadrant
 1. *begin*: node p move from one location to another location [Q0 to Q1 or Q1 to Q2 etc.] within the quadrant
 2. *move*: movement of nodes within the quadrant. After movement the information of node (radius r and quadrant number) is changed in location server of location server region
 3. *Protocol*: for communication any one of the protocols is used (e.g. DSDV, AODV, DSR, LAR etc)
 4. *Store*: New information (radius r and quadrant number) is stored into the location server
 5. *Repeat*: [step 1 to step 2] for new moving nodes

V. CONCLUSION

In this paper I propose a position based routing scheme called "Concentric Circular Quadrant Routing Scheme (CCQRS)". In this paper we use single location server region for location location update and location query. This strategy reduces the cost associated with location update and location query. This strategy reduces the cost associated with location update and location query.

VI. REFERENCES

- [1] Denin Li Chenwen wang Jian Fang, Jie Zhon, Jiaccum Wang "Reliable Backbone based location management scheme for mobile ad hoc network", 2010 IEEE
- [2] Kausik Majumdar, Subir kumar sarkar "Multilevel Location Information based routing for mobile ad hoc network."
- [3] Shyr-Kuen Chen, Tay-you Chen, Pi-Cheng wang "Spatial aware location for mobile ad hoc network", 2009 IEEE.
- [4] GS Tomar, RS Tomar, "Position based routing for mobile ad hoc network" 2008 IEEE.
- [5] Strategies for data location management in mobile ad hoc network", 2005 IEEE.
- [6] Mujtaba khambatti and Sarath Akkineni, "Location management in peer to peer mobile ad hoc network", May 2002 IEEE.
- [7] Jipeng Zhou, Zhengjun Lu, Jianzhu Lu, Shuqiang Huang "Location Service for mobile ad hoc networks with Holes", 2010 IEEE.
- [8] D.P. Agrawal and Q.A. Zeng, "Wireless and Mobile Systems," Second Edition, Thomson, 2006.
- [9] L. Blazevic, J.Y. La Boudec, and S. Giordano, "A location-based routing method for mobile ad-hoc Networks," IEEE Transactions on Mobile Computing, Vol. 4, No. 2, March/April 2005, pp. 97-110
- [10] Mobile Ad Hoc Networks (MANET) Charter. Work in progress. <http://www.ietf.org/html.charters/manet-charter.html>, 1998.

AUTHORS PROFILE



Authors1: He is pursuing Master of Engg. (M.E.) in Computer Science and Engg. From Shri Vaishnav Institute of Technology and Science Indore RGPV university Bhopal. Submitted Project Title "Location Management in MANET". Completed B.E.(Computer Science) from R.G.P.V., Bhopal.



Author2: Working as an Assistant Professor (CSE dept.), Shri Vaishnav Institute of Technology and Science Indore. M.E. from I.E.T. DAVV University.

Comparative Analysis of Techniques for Eliminating Spam in Fax over IP

Manju Jose
Research Scholar,
Mother Teresa Women's University
Kodaikanal, India
manjusaje@gmail.com

Dr. S.K. Srivatsa
Senior Professor,
St. Joseph College of Engineering,
Chennai, India.
profsks@rediffmail.com

Abstract— Fax over PSTN line suffered heavily in terms of high costs, call drops, improper imaging, corrupt messages etc. Fax over IP (FoIP) came as substitute to eliminate costly connection and transmission fees. With growing usage of FoIP, the spam is also rising and ruining business prospects. The approaches from the point of view of regulations and technologies to curb spam have been examined. Three technologies have also been compared to suggest most suitable anti-spam technology.

Keywords- Fax over IP (FoIP), S/Fax, SharePoint, Spam, Filtering Spam Fax, Digital signature

I. INTRODUCTION

Facsimile or fax over the Public Switched Telephone Network (PSTN) line revolutionized the electronic transmission of documents and made Telex obsolete. A fax machine traditionally is an electronic device having scanner, modem and a printer inbuilt. Fax machine transmits data in pulses through a PSTN line to a recipient using a compatible fax machine. The recipient fax machine transforms these pulses into images and prints the same on fax paper. The traditional method reserves its usage for PSTN line, and only one fax can be sent or received at a time.

Fax over PSTN line suffered heavily in terms of high costs, call drops, improper imaging, corrupt messages etc. There was a need to transform the way traditional fax systems worked. Instead of using PSTN line, an idea of using IP was experimented with. This experiment was found to be extremely successful as it cut down the costs and made faxing easier for users. Since fax over IP (FoIP) transmits data over an already established network, it eliminates costly connection and transmission fees.

Initially, fax over IP suffered from lack of quality as well as efficiency and was not being considered as an ideal alternative to traditional fax systems. These issues plagued fax over IP system as it didn't have a technology of its own. Initially, it used Voice over IP (VoIP) as its base technology. However,

this technology has improved over the years and now has elaborate standards and mechanisms for faxing over IP. Reducing cost is one of the most important reasons for growing importance of fax over IP. Fax over IP can be almost free in some cases. Also, fax over IP doesn't demand expensive additional equipment because the existing fax machine can be used for the purpose [16], [17].

II. FAX OVER IP

Consolidation of data and communication networks through IP is serving the objectives of reducing IT infrastructure costs while managing data and communication applications efficiently [19]. Communication technologies getting standardized through IP are causing an overlap between network applications based on traditional communications backbone and future IP environments. Organizations today are experiencing challenges in understanding their existing network applications. However, they can take advantages of new IP-based approaches to data communications including fax over IP (FoIP).

Most organizations are willing to implement VoIP to support their voice based operations. These organizations also need to provide reliable faxing capabilities to their employees. It makes commercial sense as well to utilize the existing VoIP technologies for supporting fax operations as well. There are huge potential benefits for organizations consolidating fax with voice systems through unified messaging applications. Initially, fax systems have utilized reliable PSTN networks and have been backed by reliable T.30 fax protocol to establish and maintain communication between two fax devices. Now FoIP has new standards developed specifically for FoIP making it possible for fax transmissions to utilize the Internet. Noticing this trend, fax manufacturers have started manufacturing fax equipments that can support transmission over Internet as well [19].

III. TECHNOLOGIES

FoIP is supported by two technologies such as "store-and-forward" (T.37) and "realtime" (T.38). These technologies

utilize the standard T.30 fax definition to recognize data being transferred and to ensure compatibility with existing fax devices. These technologies differ from each other in methods of delivery and confirmation receipts. Real-time FoIP, based on the International Telecommunications Union (ITU) standard T.38, elaborates the necessary technical features for transferring fax in real-time mode between two standard Group 3 facsimile terminals over Internet or other IP based networks. T.38 is generally preferred as FoIP protocol because of its ability to align with the technologies of faxes over PSTN.

T.30 handles IP fax transmission through like a standard fax call facilitating an end-to-end communication. With T.38, sending and receiving fax is similar to fax handling of non-FoIP fax devices. This involves establishing a session, sending and verifying the transmission of one or more pages and finally completing the session with positive confirmations from both sides. FoIP-enabled transmission is different as the server traverses first part of the communication to the network on IP technology rather than the PSTN. The session can use T.38 for transmission if the partner device is directly addressable on the same network. The IP switch converts T.38 packets to standard T.30 packets over PSTN if the devices are separated by the telephone line.

IP fax mechanism has two approaches namely boardless approach and the boarded approach for an Ethernet connection. Boardless system makes the fax works through DSPs (digital signal processors) located in IP routers. DSP manages transmission and conversion of T.38 packets to T.30 packets. Since there is greater awareness among organizations over implementation of VoIP, there are lesser chances that organizations will go for FoIP before VoIP.

Rather organizations are likely to utilize the already existing VoIP infrastructure to minimize the expense of adding FoIP to the same infrastructure. Also, VoIP will offer the required support already included in its routing infrastructure. Organizations need to ensure that investments in VoIP infrastructure are justified in including support for FoIP as well.

Organizations adopting an IP environment need to enhance the benefits of their existing fax technologies by enabling them to support FoIP communications.

A research [19] surveyed more than 500 fax users and potential users on their FoIP needs and expectations. This research identified the following most sought after benefits while implementing FoIP technologies:

1. Savings in Total Cost of Ownership (TCO) due to network consolidation
2. Ability to push consistent fax solution throughout the entire network including remote locations.
3. Improved IT management
4. Device/application integration

5. Least Cost Routing
6. Better utilization of new IP equipment
7. Eliminate the fax boards

VoIP networks are increasingly entering private and public organizations with IP Telephony technology. These organizations would naturally want to leverage the value and convenience of single IP communications network. Standard VoIP Codecs have been designed for voice conversations. These Codecs allow certain amount of latency and packet loss, which can still be accepted in a voice conversation.

However, faxes cannot afford to accept even small amounts of latency or any packet loss, rendering standard VoIP Codecs unacceptable for reliable faxing.

This limitation forces organizations to retain analog lines on PBX, or deploy expensive fax boards to manage their fax traffic. These extra expenses affect the ROI negatively that organizations expect from their VoIP network investment. This scenario also requires organizations to maintain legacy communication equipments with their modern IP network infrastructure.

FoIP can utilize TCP/IP standards and technologies to connect to the closest PSTN access to send and receive faxes. FoIP, as a public standard, is supported by most of the vendors' VoIP gateways. Rather than connecting a fax card to the PSTN, FoIP device can connect directly to the T.38 supported VoIP gateway. This principle is applicable on both inbound and outbound faxes [20]. eFax uses the store-and-forward capabilities as per T.37 standard to enable sending of faxes as emails. This obviously doesn't happen in real-time, which is the expected legal standing associated with fax. The fax is forwarded as email to an email server and then transmitted as an email attachment to a fax device for physical faxing.

This method cuts IT support requirements for fax, but introduces limitations such as higher costs. Other providers use their VoIP networks to transmit fax, but this method requires an Analog Terminal Adaptor (ATA) hardware device into which a traditional fax machine must be plugged. These services are not enabling the users to fax straight to a receiver desktop.

These limitations force organizations to not to use their existing VoIP networks for faxing. These organizations have limitations in realizing that VoIP capabilities added to their infrastructure also have added FoIP capabilities, without additional effort or expense [21].

There has been growing consensus regarding using existing VoIP networks to support faxing to save efforts and cost. Still, notable differences between VoIP and FoIP need to be understood by the organizations before taking a decision.

For organizations that have already implemented VoIP technologies, adding FoIP capabilities on the top of VoIP infrastructure would make perfect business sense.

However, for organizations that are yet to decide on type of IP

capabilities, they would scale up to, a thorough comparison between VoIP and FoIP is definitely required.

FoIP is less tolerant to delays in comparison to VoIP. As delay in the absence of keep-alive mechanisms may force a session to drop. However, VoIP would absorb such network impairments and would force delayed browsing experience to the users. FoIP is also less tolerant towards packet errors or losses in comparison to VoIP. For communication reliability, FoIP may require an error-correcting but delay-adding protocol such as TCP to ensure delivery confirmation or repeat requests [22]. Considering the tighter integration of FoIP with real-time transmission and reliability, FoIP is treated as first possible alternative by the organizations.

IV. THE GROWTH STORY SO FAR

A research report [2] highlighted that fax revenues would continue to grow from \$270 million in 2005 to \$400 million in 2010, achieving a remarkable compound annual growth rate (CAGR) of 8.2%. This report estimated that fax over IP sales would grow by a 50.7% CAGR to \$245 million in 2010 and non fax over IP fax sales would decrease to \$155 million in 2010. This report further added that fax over IP systems would dominate the market by 2010 with integration of VoIP.

Another research report [4] highlighted that more than 100 billion fax pages are transmitted around the globe on a yearly basis despite efforts to become a paperless society. This research report also highlighted that fax over IP markets would continue to grow rapidly during the five-year forecast period unfazed by economic recession. This report further highlighted that fax has certain advantages over e-mail as it largely remains compatible across various devices and systems, retains the format of complex documents and finally sends documents in non-editable format so that the recipients do not modify them.

A research report [3] on fax messaging markets observed that patent law situation with respect to fax over IP has undergone radical changes. This report further added that fax over IP would continue to enjoy growth rates that were earlier diminished due to economic downturn. This report concluded with fax over IP market projections of \$ 1.575 billion in 2013. A research report [5] on computer based fax markets for the period of 2009-2014 and observed that usage of IP has contributed towards tremendous growth of fax over IP. This report highlighted phenomenon of multi function peripherals (MFPs) and integration of business computing systems. The MFPs would account for 32% of revenues produced through fax over IP industry. This integration has also fuelled demand for fax over IP services. This report forecasted that desktop faxing would account for 70% share of fax systems, out of which 42.5% share would be attributed to unified messaging.

V. SPAM IN FAX: THE INSIDER'S PERSPECTIVE

An article [1] argued that fax over IP systems have brought

big relief to business users with numerous advantages associated with it such as compatibility, retention and

integrity. This article also highlighted the challenges of spam or junk fax that were ruining the business prospects of fax over IP systems.

The perspectives of spam fax differ for the recipients and organizations dealing in fax marketing [14]. The costs attached with spam are high and fax spammers are not authorized to use fax paper and toner of unwilling recipients to send sales pitches. Further, fax spammers shouldn't be allowed to tie up bandwidth, computing power, and storage of unintended recipients as they haven't paid for it [18]. With aggressive marketing pitches over fax machines, email boxes and cell phones, the list of offenses that deserve the spam handle is also growing [10].

VI. THE FIGHT BACK

A new lobby of anti spam activists has decided take on fax spammers and these activists have been fighting against unsolicited fax messages through claims of damages and advocating for a law on spam faxes.

A suit [9] filed to get junk or spam faxes banned was different in approach as it didn't claim damages. Rather, it helped in establishing a junk fax ruling under which sending of junk faxes was considered a criminal act.

A prominent case [13] used junk fax law to sue violator over email spam and sought damages proving the ground comprehensively. This case argued that computer was also acting as fax machine and that email was really no different from fax over IP. This case further argued that received spam email using a phone line connection and printed message demonstrate suitability of trial of this case under junk fax law.

An article [8] cited doubts in the opinions regarding willingness of courts to apply junk fax law, as most of the existing state laws on spam are weak and counterproductive.

VII. SPAM FAX: CURBING THE MENACE

An article [14] suggested certain measures to prevent spam in fax over IP as per the provisions of the Junk Fax Prevention Act [15]. The recipient can contact the sender directly and express his/her unwillingness to receive spam fax messages. The recipient can also choose "opt out" option to block his/her fax number from fax mailing lists. The recipient also has a choice to approach a regulator and file a complaint or even claim damages.

A suggestion was made regarding integrating FoIP with Microsoft SharePoint to eliminate spam and security risks. This integration can ensure the right movement of documents and junk or spam is filtered out [1]. This article further elaborated that FoIP, when integrated with SharePoint, provides efficient inbound routing of fax into SharePoint sites. This feature enables the users to search content, metadata tags, and optical character recognition (OCR) features of fax

documents to differentiate spam from necessary documents and apply rule based routing. The users also have the provision of triggering an automated SharePoint work flow process to ensure that fax documents go to the right folders or destinations while filtering junk out of the system.

Another article [7] criticized the spam filters for their application in content sorting only. Moreover, most of the spam filters sort the spam after accepting the messages. High volume of spam forbids the users from receiving it and all messages classified as spam are automatically deleted. This causes difficulties to genuine senders as their messages are deleted without any reason being assigned. Using spam filter on protocol level would reject the spam and assign reason for doing so. Such spam filters are rare. This article advocated the use of pdf format over spam filters to counter spam faxes. A suggestion was made regarding using secure fax (S/FAX) over IP to generate secure mails in pdf format containing sender's name, address, and digital signature. This pdf can also be given password protection, if required. Upon receipt, the pdf reader can verify the integrity of the document as well as the digital signature and allow the receiver to either view it or discard as spam.

An article [11] suggested charging fax spammer from \$500 to \$1500 for every unsolicited fax they send. The article added that this amount could become significant if fax spammers send many faxes. The recipient needs to keep track of all unsolicited faxes with date & time stamps to be used for claims.

A technology [6] has also been proposed regarding spam fax filter that can transform a rasterized form of a fax image into non-rasterized forms. Non-rasterized forms of the fax image can then be checked against spam fax characteristics, or characteristics of fax images not known as spam. The fax image is declared as spam if it tests favorably to at least one of the known characteristics of spam fax. In case, the fax image doesn't conform to any of the characteristics of spam fax, it can be declared as non spam.

Another article [12] suggested complaining to regulator and highlighting the matter to draw mass attention to curb the fax spam menace.

VIII. COMPARATIVE ANALYSIS

A comparative analysis as given in table.I involving three approaches [1], [6], [7] has been done to understand the unique characteristics associated with each approach.

While comparing, the key indicators such as disposal of spam faxes, system overload, response time etc were kept in consideration. The researcher had attached higher significance to anti fax spam technology offering pre disposal of spam faxes as this would largely save the network from getting flooded with spam faxes resulting in higher load on resources.

TABLE I. Comparative Analysis

Technologies	Integration of FoIP with SharePoint	S/FAX over IP (PDF Format)	Spam Fax Filter (Non- rasterized Format)
Mechanism	Inbound rule base routing of fax into SharePoint sites	Use of S/FAX over IP to generate secure mails in pdf	Conversion of rasterized form of fax image into non-rasterized form
Filtering / Segregation	Searching content, metadata tags and OCR features of fax documents to filter spam from necessary fax documents	Verifying integrity and digital signature of pdf documents to allow fax messages to enter	Verifying non-rasterized forms of the fax image against spam fax characteristics to filter spam from necessary fax documents.
Load on Network	High	Low	High
Disposal	After fax enters the system	Before fax enters the system	After fax enters the system
Response Time	High	Low	High

IX. CONCLUSION

S/FAX over IP approach proposed by Engelbert is conclusively better than rest of the approaches, as this approach does not allow the spam fax to get into system. Rather, this approach proposes to use spam filter to accept or discard fax messages subject to verification of integrity and digital signature of incoming fax messages. Also, this approach proposes to assign reasons for rejection of fax messages enabling the senders to receive the status of their messages.

X. LIMITATIONS

The researcher experienced certain limitations while conducting this research as there is no benchmark available, which could be taken into consideration for comparing the technical approaches to prevent spam in fax over IP. The researcher interacted with fax users to understand their expectations from these spam prevention approaches to identify the key indicators for comparison. The researcher attached high significance to pre disposal of spam faxes as this would prevent the spam faxes from entering into the system. The significant factors may be different for different users.

XI. FUTURE RESEARCH WORK

The researcher is willing to work and explore this area more in order to identify best technologies and approaches to prevent spam in fax over IP. The researcher also intends to bring more technologies and approaches under the ambit of comparisons. A detailed research can be initiated on the basis of already covered issues and suitable statistical parameters can be used for rigorous technical comparisons. Developing a globally acceptable technology or approach to prevent spam in fax over

IP is also a possibility in future.

REFERENCES

- [1] Campbell, S. (2010), The Many Benefits of FoIP and Microsoft PowerPoint, <http://www.tmcnet.com/channels/foip/articles/77591-many-benefits-foip-microsoft-sharepoint.htm>.
- [2] Davidson Consulting (2006), Computer-Based Fax Markets, 2005–2010, December 2006
- [3] Davidson Consulting (2009), Fax Messaging Markets, 2008–2013, July 2009
- [4] Davidson Consulting (2009), FoIP Server Markets, 2007–2013, April 2009
- [5] Davidson Consulting (2010), Computer-Based Fax Markets, 2009–2014, April 2010
- [6] El-gazzar, Amin et al. (2007), Spam fax filter, United States Patent 7184160.
- [7] Engelbert, S. (2005), How to send a S/FAX over IP, <http://www.aloaha.com/wi-software-en/sfax-ip.php>.
- [8] Festa, P. (2002), Newsmaker: A cybersage speaks his mind, CNET News, http://news.cnet.com/A-cybersage-speaks-his-mind/2008-1082_3-958576.html?tag=mncol;txt.
- [9] Festa, P. (2003), Spam law a matter of fax? CNET News, <http://news.cnet.com/2100-1028-994076.html>.
- [10] Hansen, E. and Olsen, S. (2002), Spam: It's more than bulk e-mail, <http://news.cnet.com/2100-1023-961134.html>.
- [11] Miller, N. (2010), Stop Fax Spam For Good, <http://ezinearticles.com/?Stop-Fax-Spam-For-Good&id=3577712>.
- [12] Nefer, B. (2010), How Do I Cancel Spam Faxes? http://www.ehow.com/how_6885111_do-cancel-spam-faxes_.html.
- [13] Seymour, J. (2003), Michigan Man Uses Junk FAX Law to Sue Sears Over Spam, http://www.linxnet.com/misc/spam/mi_spamsuit.html.
- [14] Stevens, P. (2006), Spam Fax Facts, <http://faxing-service-review.toptenreviews.com/spam-fax-facts.html>.
- [15] The Telephone Consumer Protection Act (2005), http://www.junkfaxes.org/federal_law.htm.
- [16] Unuth, N. (2010), IP Faxing: A Cheaper Way of Sending Fax, <http://voip.about.com/od/faxingoverip/a/IPFaxingIntro.htm>.
- [17] Unuth, N. (2010), Reasons for Using Faxing over IP, <http://voip.about.com/od/faxingoverip/tp/AdvantagesofIPFaxing.htm>.
- [18] Warner, J. (2003), How much does spam really cost? The Miami Herald, March 10, 2003.
- [19] The Essential Guide to Fax Server Software, Windows ITPro eBooks, May 2009.
- [20] Choosing between Traditional and FoIP/VoIP Faxing, Sagem-Interstar Group, July 2007.
- [21] Chemicoff, D., The Essential Guide to Fax over IP, Windows ITPro, April 2006.
- [22] Berkowitz, C. H. (2006), Save money by using fax over IP, Computerworld.

Resource Estimation and Reservation for Handoff Calls in Wireless Mobile Networks

K.Venkatachalam (Corresponding author)

Department of ECE,
Velalar College of Engineering and Technology,
Thindal post, Erode – 638012, Tamilnadu, India.
venki_kv@yahoo.com

Dr.P.Balasubramanie

Department of CSE
Kongu Engineering College, Perundurai,
Erode, Tamilnadu, India
pbalu_20032001@yahoo.co.in

Abstract— The main aim of future wireless multimedia networks is to provide sufficient amount of resource to a multimedia call. Reserving required amount of resource in advance to a future new and handoff call is better than rejecting a call at neck of the moment due to insufficient resource at a particular time. This paper presents an efficient handoff resource management strategy by considering the future resource demands of a wireless multimedia call. Here a novel wiener based resource estimation and reservation method is adopted to estimate the instantaneous resource demands of a mobile user. Cell segmentation technique is introduced and utilized to predict the resource demands more accurately in a real time manner. The performance result shows this synergy of resource reservation using pilot sensing and cell segmentation have been decrease the call dropping probability of the handoff calls and increases the micro and pico cellular system performance in real time environments.

Keywords—component; Handoff(HO), Resource Estimation(RE), Resource Reservation(RR), Cell Segmentation (CS), Call Dropping Probability (CDP), Wireless Multimedia Networks (WMN).

I. INTRODUCTION

Resource allocation is done to ensure an efficient use of resources in the wireless network. Radio resource allocation in cellular mobile system focuses mainly to improve the user admission capability and protecting the connection continuity. Handoff (HO) is an operation in which Mobile Unit (MU) communicating with one wireless Base Transceiver Station (BTS) is transferred to another base transceiver station during a call. A wireless mobile call in progress could be forced to abort during handoff, if it cannot be allocated sufficient amount of resources in the new wireless cell. A cell is the radio coverage area of a wireless base transceiver station. Present wireless cellular systems are employed with mobile assisted soft handoff technique for handoff operation. Handoffs are critical in cellular mobile system because neighboring cells are always using a disjoint subset of frequency bands. Hence negotiations must take place between mobile units, the current serving base transceiver station and the next potential base transceiver station.

Reserving resources for future handoff calls and new calls is an effective way to reduce the handoff call dropping and new call blocking probability. Predicting and reserving resources for future calls can be classified into two types. They are local and collaborative methods[1]-[5]. Existing collaborative and local methods for resource reservation requires each base transceiver station to gather real time information on the behaviors of mobile units in neighboring cells. Such information may include how many users are expected to be handoff and service class of multimedia call in the neighboring cells at a given time. Local methods [6]-[8] assumes that every call requires the same bandwidth, the call arrival process is poisson, and the call holding time and a particular call channel holding time in each cell is exponentially distributed. Service class of a multimedia call mainly deals with how much amount of resource required for each call request. In the real time environments gathering the above information in a very short duration is very difficult one.

The mobility-dependent predictive resource reservation (MDPRR) scheme and an admission control scheme are proposed in [9] based on common handoff procedure to provide flexible usage of scarce resource in mobile multimedia wireless networks. NPS(neural-network prediction scheme) is proposed in [10] to provide high accurate location prediction of a MH (mobile host) in wireless networks. In order to avoid too early or over reservation resulting in a waste of resources, a three-times resource-reservation scheme (TTRR) is also proposed. The work in [11] is based on application of multi-input-multi-output (MIMO) multiplicative autoregressive-integrated-moving average (ARIMA) $(p,d,q) \times (P,D,Q)_s$ models fitted to the traffic data measured in the considered cell itself and on the new call admission control (CAC) algorithm that simultaneously maximizes the system throughput while keeping the handoff call dropping probability (CDP) below the targeted value. The mobility-aided adaptive resource reservation (MARR) with admission control (AC) based on cell division, to provide better usage of scarce resource in wireless multimedia networks is proposed in [12]. The effect of pre-reservation area size on handoff performance in wireless cellular networks are discussed in [13]. It shows that if the reserved channels are strictly mapped to the MSs that made the corresponding reservations, as we increase the pre-reservation area size, the system performance (in terms of the probability

that the handoff calls are dropped) improves initially. The optimal pre-reservation area size is closely related to the traffic load of the network and the MSs' mobility pattern (moving speed).

II. LIMITATIONS OF AVAILABLE METHODS

Existing local and collaborative methods for predicting and reserving resources for future handoff calls and new calls are not much suitable for wireless multimedia networks. This is because of the following reasons.

The amount of resource required to successfully handoff a call may vary over a wide range in a multimedia wireless networks. For example, data and video application calls may require different service quality levels and consequently require different amount of resources in order to ensure a successful handoff. Wireless networks are often consist of large number of micro and pico cells (i.e., very small radius cells). In such networks, handoff becomes more frequent, handoff call arrivals may be non-poisson and non-stationary for extended periods of time, and a handoff call channel holding time distribution inside each cell can be arbitrary. Even in macro cellular networks, handoff call arrivals may often be non-poisson and non-stationary for extended periods of time. For example, handoff call arrival rates will vary with the number of mobile users, user mobility pattern and network configuration.

Speed of mobile units may vary widely and mobile users may stay in a particular micro or pico cell for very short time periods. Hence gathering real-time information on current status and behaviors of mobile units in other cells and communicating such information among base transceiver stations in a timely fashion will increases the system complexity and cost.

The limitations of existing methods caused primarily by, they do not model the resource demands of handoff calls and new calls directly. In a real multimedia wireless networks, number of factors can impact the resource demands of future handoff calls and new calls. They include cell sizes, network configuration, number of mobile units in each cell, speed and mobility pattern of mobile units, types of services supported in each cell, types of services used by each mobile unit at any given time, arrival processes of new and handoff calls, call and channel holding times, etc. These factors often have a complex correlation and the set of the factors often changes over time. Consequently, modeling these factors can be difficult, especially when only local information is available. This is primarily why most existing collaborative and local methods can only handle poisson and stationary call arrivals, and requires each radio channel to have the same capacity.

In this work a new class of Cell Segmentation (CS) based new call and handoff call resource estimation and reservation method is proposed. This overcomes some of the critical limitations of existing methods by modeling the instantaneous amount of resource demands directly. The proposed RER method uses pilot sensing method to gather information. This method perform well for new and handoff call arrival processes are non-poisson and non-stationary and each call requests an arbitrary amount of resources i.e. limit allowed by the network

and has non-exponentially distributed call and channel holding times.

III. KEY FEATURES OF PROPOSED RESOURCE ESTIMATION AND RESERVATION (RER) METHOD

Here a new class of dynamic resource estimation and reservation method for supporting multimedia call is proposed. The proposed RER method has the following properties.

- *Localized prediction*: Each base transceiver station uses local available information from neighboring cells to determine dynamically how much resource should be reserved for future handoff calls and new calls. It communicate with other base transceiver stations for resource reservation decision, depends upon Predetermined Time Interval (PTI).
- *Modeling instantaneous demands directly*: The proposed method models the instantaneous values of resource demands directly by using cell segmentation. It also enables to predict instantaneous and average future demands, while other existing methods can typically predict only average demands.
- *Multimedia call resource estimation and reservation*: The proposed method estimates the future resource demands of each individual service class of multimedia call directly. It can also estimate the total amount of resource required for handoff calls and new calls of all service classes of a multimedia call.
- *Simplicity*: The proposed method is much simpler to implement in real time and existing networks.

IV. PROPOSED RESOURCE ESTIMATION AND RESERVATION METHOD

The proposed resource reservation method for handoff calls is shown in Figure 1 and is a self explanatory one. When a mobile unit is approaching towards the cell boundary, its position and velocity are monitored. By using this, its remaining time in the current cell is calculated. Once this time falls below the threshold value called Resource Reservation Interval (RRI), then an new channel reservation request is sent to the test or target cell.

If there are free or ideal channels in the target cell, then one channel is immediately reserved. At the same time, the channel is locked and temporarily disabled for other usage in the target cell. If the target cell has no free channels, then the reservation request waits for predetermined time interval. When a channel is released in the target cell within PTI, then that free channel is assigned for demand request. If there is no reservation request arrives then the released channel is remains free until the next channel request arrives. When a mobile unit ends its call connection in the current cell, but moving towards the target cell, in this case, a reservation cancellation request is forwarded to the target cell. Upon receiving the cancellation request, the target cell releases the locked channel or clears the reservation request.

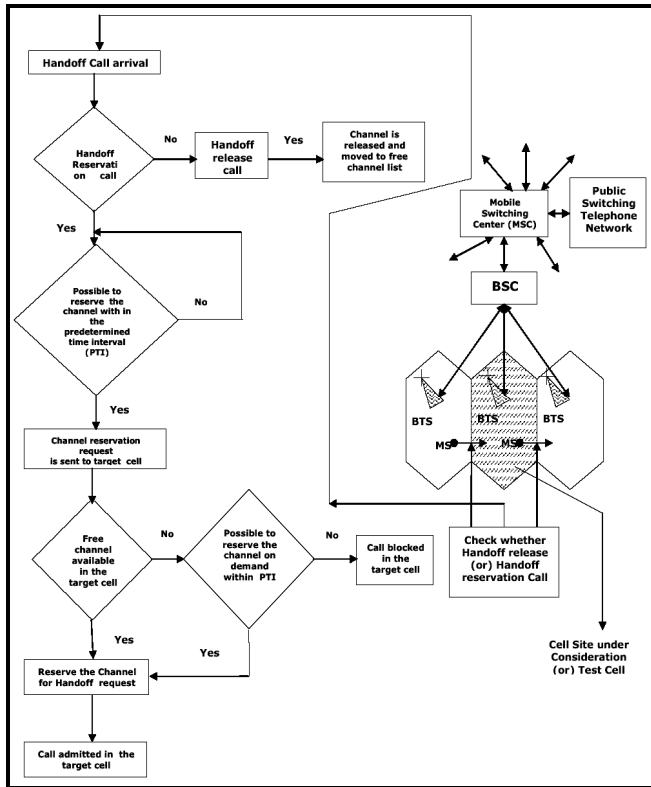


Figure 1. Proposed Handoff Call Resource Reservation Method

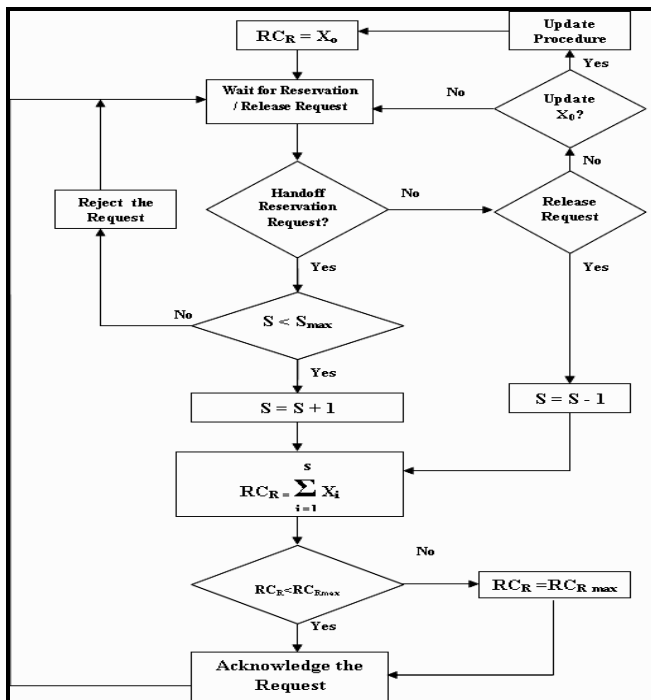


Figure 2. Flow Chart for Resource Reservation / Release Operations

RC_R - Reserved Resource Capacity
 RC_{Rmax} - Maximum Reserved Resource Capacity
 S - Number of reservations

When the mobile unit enters the target cell, handoff will be successful only if a channel has been reserved to take it over, or blocked if its reservation request is not yet processed. In the former case, the mobile unit continues its call, on the new channel until leaving or call completion, while in the later case, the call is forced into termination. A new call is accepted only if a free channel exists upon its arrival. Otherwise it is blocked and cleared from the system i.e. in the channel servers. For new calls, there is no need of resource reservation. Once free channels are available, and then connection is established. Otherwise the caller has to wait until the availability of free channels in the current cell. But for handoff calls, the resource reservation in advance is a mandatory one.

V. FLOWCHART FOR RESOURCE RESERVATION/ RELEASE OPERATIONS

The flow chart for channel reservation /release operation is shown in Figure 2.

- The reserved resource capacity $RC_R(S)$ is initially set at X_0 and Base Station Controller (BSC) waits for reservation request.
- When a channel reservation is requested by mobile unit, the associated BSC accept the request, if the number of reservations ' S ' in BSC is smaller than the predetermined maximum value of S , S_{max} .
- In the case of acceptance of the reservation request, the BSC increases ' S ' by one and increases $RC_R(S)$ by X_S . Which can be properly set at a different value for each ' S ', if the reserved capacity $RC_R(S)$ is less than the RC_{Rmax} . Otherwise the $RC_R(S)$ is set at RC_{Rmax} .
- When the release of the reserved channel capacity is requested, the BSC decreases ' S ' by one and $RC_R(S)$ by X_{S+1} , if $RC_R(S)$ is less than the RC_{Rmax} , otherwise $RC_R(S)$ remains at RC_{Rmax} .

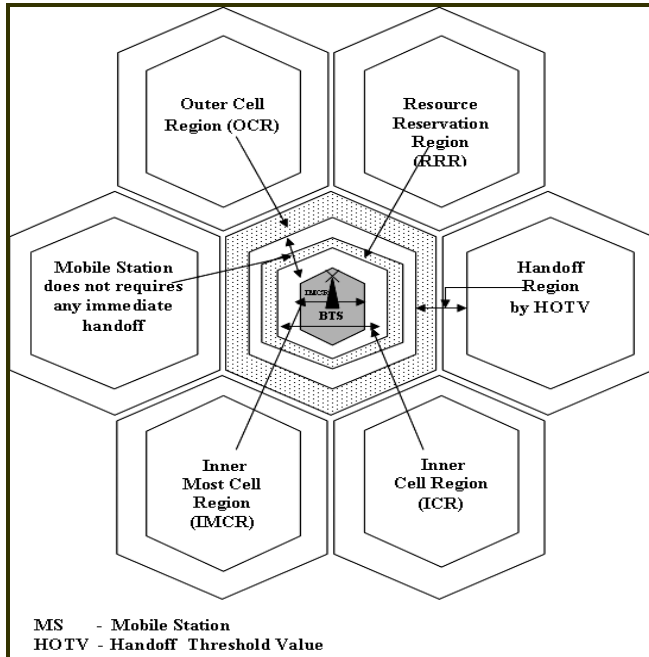
This method of pilot sensing reservation mechanism reduces the unnecessary blocking of new calls and dropping of the hand of calls. Since the system capacity depends on the new call blocking and handoff call droppings. The system capacity is limited by new call blocking if the new call blocking probability is higher than the weighted sum of the handoff call failure probability. If the weighted sum of the handoff call failure probability is higher than the new call blocking probability, the excursive handoff call failure probability limits the system capacity.

The system capacity is maximum, when the new call blocking probability is equal to the weighted sum of the handoff dropping probability. To keep a balance between the new call blocking and the handoff call dropping probability, this method controls the size of the minimum reservation capacity X_0 by counting the number of new call blocking and handoff call droppings. For resource reservation, new calls and handoff calls are taken into account, and the Markov model shown in Figure 4 can be used to reduce the computational complexity.

VI. CELL SEGMENTATION FOR RESOURCE RESERVATION / RELEASE OPERATIONS

For effective resource reservation process as well as the handoff process, the cell area is divided in to the following regions. It is shown in Figure 3. They are

- The Inner Cell Region (ICR)
- Inner Most Cell Region (IMCR)
- The Resource Reservation Region (RRR)
- Handoff region by Handoff Threshold Value (HOTV).
- BHOTV – Below Handoff Threshold Value
- The Outer Cell Region (OCR)



	MU Resource Release conditions	MU Resource Reservation Conditions
1.	IMCR to RRR - does not require any immediate handoff	RRR to IMCR - Does not require any reservation
2.	RRR to outer cell region requires the channel release in the current cell	MU entering from neighboring cell to test cell outer region then test cell requires immediate handoff reservation

Figure 3. Cell Segmentation for Resource Reservation/Release Conditions

A Communicating MU requires Resource Reservation for the following conditions

- When the mobile unit moves from RRR to IMCR then it does not require any immediate resource reservation.
- When the mobile unit enters from the neighboring cell outer boundary i.e OCR to the test cell outer region, then immediate reservation is required in the test cell.

A communicating mobile unit does not require any further resource reservation for the following conditions.

- The call of a communicating mobile unit is terminated with in the RRR.
- The mobile call is terminated with in the IMCR after the caller moves from RRR to OCR.

A Communicating mobile unit in RRR region requires another reservation for the following conditions.

- A communicating mobile unit moves from RRR to IMCR and moves back into the RRR region again.
- A communicating mobile unit moves from the RRR to another RRR.

When the mobile unit moves from RRR to HOTV i.e. handoff threshold value, then the threshold value of the received signal in the MU decreases and handoff occurs. This handoff is a handoff release process and the channel is kept in the Base Station Controller (BSC) or Mobile Switching Centre(MSC) pool for allocating this channel into other channel reservation requests. When a new call arrives in an RRR region, then it requires an immediate channel reservation if it is not blocked.

VII. MARKOV MODEL

In the Markov process discussed by Lee[14] and Taha et al[15], future value is independent of the past values, given in the present value. i.e. It can models the future, depends only upon the present state. The Wiener process and Poisson process are Markov process. Since, both have the properties of independent increment, in a continuous time space. If the call arrival at a particular time interval is minimum in numbers, then poisson process is suggested. If the call arrival at a particular interval of time is maximum in numbers, then Wiener process is suggested.

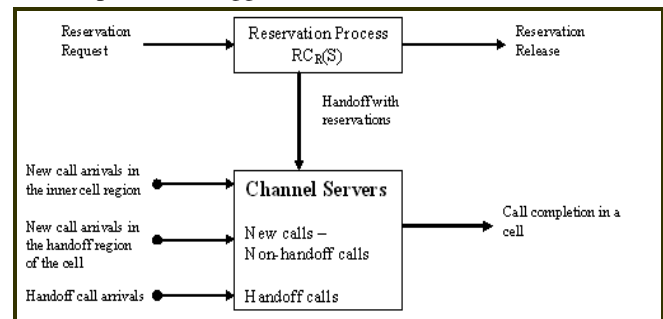


Figure 4. Markov model for new call, handoff call and resource reservation call arrivals

Here Wiener process is suggested, since in micro and pico cellular systems, call arrivals in a particular time interval is maximum in numbers. New call arrivals, handoff call arrivals and reservation request arrivals based on the Markov model is shown in Figure 4. for resource reservation conditions. New call arrivals in a handoff region are admitted only if both of the associated base transceiver stations accept, and, if they are admitted, the calls go into handoff. A handoff arrival with a channel reservation enters the system as a reservation request

arrival. When the reservation is released, it is assumed that, the release is caused by a handoff attempt. Service for a reservation request arrival is completed either if the call of the MU that requested the reservation is terminated or if the MU moves out of an RRR region.

A. Wiener Based Prediction Method

The proposed technique for resource reservation is wiener based method discussed by Taha et al[15]. This method supports multimedia calls since multimedia calls are variable bandwidth calls and it supports poisson and random arrival of calls in the network. It also supports stationary and non-stationary call arrivals in the system. In the wiener process, the present and future values are affected by large number of independent or weakly dependent factors. Since in the wireless cellular network, all calls are continuous with respect to time but discrete with respect to events in nature. The expression for estimating future demand by using present demand and present estimated demand is given by

$$\hat{D}_{t+1} = (1 - \alpha) \hat{D}_t + m \alpha D_t \quad (1)$$

where,

D_t - Observed demand of the mobile unit at time 't'

$\hat{D}_t$ - Estimated demand of the mobile unit at time 't'

$\hat{D}_{t+1}$ - Estimated demand of the mobile unit at time 't+1'

α - Smoothing factor lies between zero and one

m - Amplification factor according to the value of alpha (α)

$\Delta t = 't+1'$ - Estimation time interval - 10 minutes

When $\alpha = 0$

$\hat{D}_{t+1} = \hat{D}_t \Rightarrow$ Future estimated demand is equal to present estimated demand and,

When $\alpha = 1$

$\hat{D}_{t+1} = D_t \Rightarrow$ Future estimated demand is equal to present demand, so from $\alpha = 0.1$ to 0.9 future estimated demand is calculated.

To calculate the demand more accurately, the value of ' α ' is calculated as

$$\alpha = C \frac{Es^2}{\sigma_{t+1}^2}, \quad (2)$$

where $0 < C < 1$

In equation (2),

$Es = D_t - \hat{D}_t$ is the prediction error, and

$\sigma_{t+1}^2 = CE_s^2 + (1 - \alpha)\alpha_t$, Since ' α ' is standard

normal variable and the value should lies only between zero and one

$\sigma_t \Rightarrow$ std deviation at time 't'

$\sigma_{t+1} \Rightarrow$ std deviation at time 't+1'

The wiener estimation method has the following properties

- $\Delta R = \hat{D}_{t+1}$ is modeled as normal random variable for a given $\Delta t = t+1$. Normal distribution is justified because ' α ' is the standard normal variable and the

present and future values affected by large number of independent or weakly dependent factors.

- The values of $\Delta R = \hat{D}_{t+1}$ for any two disjoint time intervals are independent in nature.
- The value $\Delta R = \hat{D}_{t+1}$ for the given time interval $\Delta t = t+1$ is independent of starting time interval.
- $\Delta R = \hat{D}_{t+1}$ has mean zero and standard deviation $\sqrt{\Delta t}$.

VIII. PERFORMANCE EVALUATION AND CONCLUSION

The analysis of wiener based future resource estimation using present demand and present estimated demand is given in section 7.1. By varying the smoothing factor alpha, it can predict the future estimated demand more accurately. The performance of a proposed wiener based resource estimation method is shown in Figures 5 and 6. Figure 5 shows linear and Figure 6 shows the non-linear estimation of resource demands of a mobile unit in a predetermined time interval. The proposed method utilizes the cell segmentation effect for resource estimation and reservation.

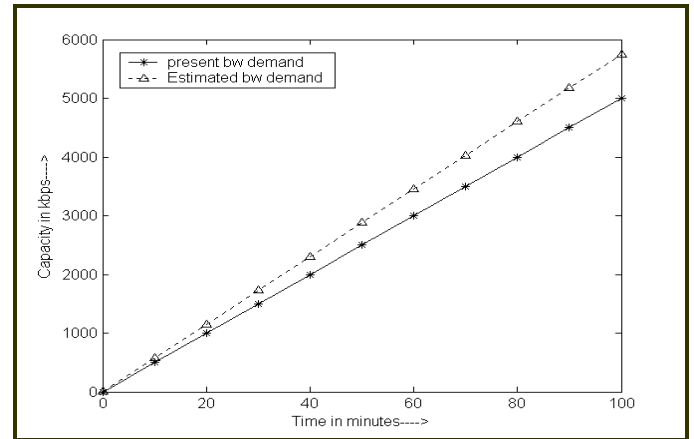


Figure 5. Resource Estimation Using Proposed Wiener Method – Linear Prediction

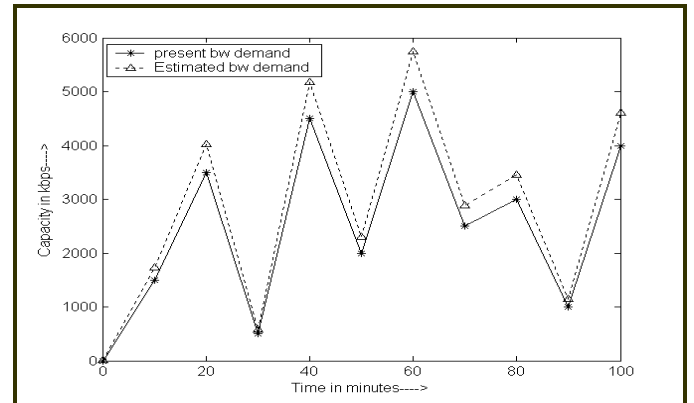


Figure 6. Resource estimation Using Proposed Wiener Method– Nonlinear Prediction

Assuming each cell can support up to 78 channels and the target Call Dropping Probability (CDP) is 5%. The call arrival rate is varied to allow a comparison of CDP between lightly loaded and heavily overloaded systems. From the simulation it is shown that, using the predictions generated by the proposed RER, based on wiener method and the existing method, the resulting CDP is comparably reduced. Utilizing the cell segmentation effect for resource estimation and resource reservation for future handoff calls effectively utilizes the available resources.

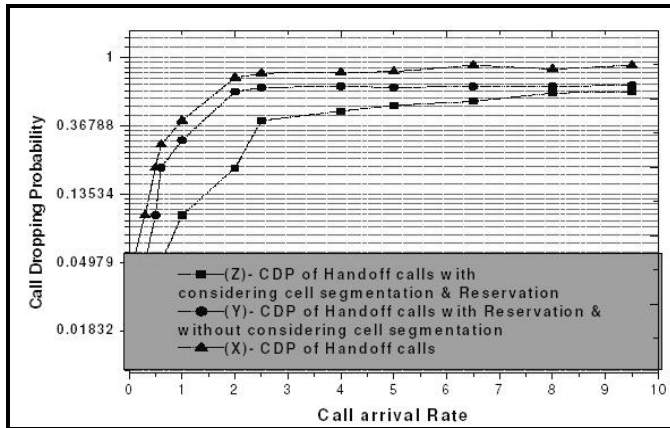


Figure 7. Handoff Call Dropping Probability Comparison between Existing and RER Method

Figure 7 shows the call dropping probability of handoff calls with and without considering the cell segmentation and resource reservation effects. In this graph, curve 'x' shows the CDP of handoff calls for existing methods of local and collaborative methods. Curve 'Y' shows the CDP of handoff calls with resource reservation and without considering the cell segmentation effect. In this case the CDP of lightly and heavily loaded system is constantly reduced by an amount of 15 percent when compared with existing methods. Curve 'Z' is the response result by utilizing the cell segmentation effect for resource reservation. Now the performance is initially improved by an amount of 35 percent in the lightly loaded system and when the call arrival rate is increased, it produces an improvement of 20 percent reduction in CDP for heavily loaded system as compared with existing methods. With considering cell segmentation for reserving the resources considerably reduces the call dropping probability of handoff calls than without cell segmentation. This synergy of resource reservation and cell segmentation effectively manage the available resources in the network and will increase the micro and pico cellular system performance in real time manner. In this work more concentration given to handoff calls because termination of an ongoing call during handoff due to insufficient resource will onset the mobile users more than the new call termination during the call initiation.

REFERENCES

- [1] P. Agrawal, D.K. Anvekar, and B. Narendran, "Channel management policies for handovers in cellular networks," Bell Labs Tech. J. Luicent Technol., Vol. 1. no. 2. 1996.

- [2] T.W. Yu and V.C.M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," IEEE J. Select. Areas Commun, Vol.15, September 1997, pp. 1208-1225.
- [3] J.S. Shuh and R.R. Rao, "A multi-rate resource control (MRRC) scheme for wireless/mobile networks," in Proc. ICC'99.
- [4] A.L. Beylot, S. Boumerdassi, and G. Pujolle, "A new prioritized handoff strategy using channel reservation in wireless PCN," in Proc. Globecom1998.
- [5] B.M. Epstein and M.Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," IEEE J. Select. Areas Commun., Vol. 18, March 2000, pp. 523-534.
- [6] X. Luo, I Thng, and W. Zhuang, "A dynamic pre-reservation scheme for handoffs with GoS guarantee in mobile networks," in Proc. IEEE Int. Symp. Computers Commun, July 1999.
- [7] L. Ortigoza-Guerrero and A.H. Aghavami, "A prioritized handoff dynamic channel allocation strategy for PCS," IEEE trans. Vehicle. Technol., vol 48, July 1999, pp. 1203-1215.
- [8] S. Kim and T.F. Znati, "Adaptive handoff channel management schemes for cellular mobile communication systems," in Proc. ICC 1999.
- [9] Li-Liann Lu, Jean-Lien C. Wu, Wei-Yeh Chen, "The study of handoff prediction schemes for resource reservation in mobile multimedia wireless networks", International Journal of Communication Systems, John Wiley and Sons Ltd. Chichester, UK, Volume 17, Issue 6, 2004, Pages: 535 - 552.
- [10] Shiang-Chun Liou, Hsuan-Chia Lu, Kuo-Hsien Yeh, "A capable location prediction and resource reservation scheme in wireless networks for multimedia," IEEE International Conference on Multimedia and Expo, 2003 vol. 3, pp. 577-580.
- [11] Rozic Nikola, Kandus Gorazd, "MIMO ARIMA models for handoff resource reservation in multimedia wireless networks", Wireless communications and mobile computing, Wiley, Chichester, vol. 4, no.5, 2004 pp. 497-512.
- [12] Li-Liann Lu, Jean-Lien C. Wu and Hung-Huan Liu, "Mobility-aided adaptive resource reservation schemes in wireless multimedia networks", Recent Advances in Wireless Networks and Systems, Elsevier Ltd, Volume 32, Issues 1-3, January-May 2006, Pages 102-117.
- [13] Zhenqiang Ye, Lap Kong Law, Srikanth V. Krishnamurthy, Zhong Xu, Suvudhean Dhirakaosal, Satish K. Tripathi and Mart Molle, "Predictive channel reservation for handoff prioritization in wireless cellular networks", Int. Journal of Computer Networks, Volume 51, Issue3, February 2007, Pages 798-822.
- [14] Lee, W.C.Y. "Mobile Communications Engineering-Theory and Applications" Second edition, Tata McGraw Hill publications, 2008.
- [15] Taha, H.A., Natarajan, A.M., Balasubramanie, P. and Tamilarasi, A. "Operations Research-An introduction", Eight edition, Pearson Education, 2008.

AUTHORS PROFILE



K.Venkatachalam graduated from Bharathiar University, Coimbatore, India, in Electronics and Communication Engineering during the year 1992. He obtained his Master Degree in Technology with Distinction & University Second Rank in Electronics & Communication Engineering - Communication System Specialization, from Pondicherry University, Pondicherry, India, during the year 2003. Now he is pursuing his Ph.D in the area of Wireless Mobile Communications at Anna University, Chennai, India. Presently he is with the department of Electronics and Communication Engineering, Velalar College of Engineering and Technology, Erode, Tamilnadu, India. He has published more than 10 Research papers in National and International Conferences & journals. His area of interest includes Computer Networking, Security Systems etc.,



Dr.P.Balasubramanie has completed his M.Phil degree in Anna University Chennai in 1990. He has Qualified for National level eligibility test conducted by Council of Scientific and Industrial Research(CSIR) and Joined as a Junior Research Fellowship(JRF) in Anna University, Chennai. He has completed his Ph.D degree in Theoretical Computer Science in 1996. He has completed 16 years of service in teaching. Currently he is Professor, Department of Computer Science & Engineering, Kongu Engineering College,

Tamilnadu, INDIA. He is the recipient of Best Staff Award consecutively for two years in Kongu Engineering College. He is also the recipient of Cognizant Technology Solutions(CTS) Best faculty award 2008 for outstanding performance. He has published more than 80 research articles in International and National Journals. He has authored 8 books with the reputed publishers. He has guided 11 part time Ph.D scholars and number of scholars are working under his guidance on various topics like image processing, data mining, networking and so on. He has organized several AICTE sponsored National seminar/ workshops.

Low Complexity Scheduling Algorithm for Multiuser MIMO System

Shailendra Mishra
Kumaon Engineering College, Dwarahat,
Uttarakhand, India
email:skmishra1@gmail.com

D.S. Chauhan
Uttarakhand Technical University,
Dehradun, Uttarakhand, India
(email:pds@gmail.com)

Abstract— Multiple-input and Multiple-output (MIMO) is one of several forms of smart antenna technology. Multiuser downlink scheduling problem with n receivers and m transmits antennas, where data from different users can be multiplexed is discussed in this paper. Scheduling Algorithm targets to satisfy user's QoS by allocating number of transmit antennas. Scheduling performance under two different types of traffic modes is also discussed: one is voice or web-browsing and the other one is for data transfer and streaming data. We have proposed scheduling algorithm for MIMO system which targets to satisfy user's QoS by allocating the number of transmit antennas.

Keywords- MIMO, SM, STC, DIV, SA, STBC, MRC, DPC, IMD

I. INTRODUCTION

The process of technological advancement has given rise to develop MIMO technology in the field of wireless communication. MIMO system also reduces the expenditure for using extra bandwidth or the transmit power expenditures and increases in throughput and range are possible at the same bandwidth. MIMO system explores the idea of multipath propagation to increase data throughput and range, or reduce bit error rates rather than attempting to eliminate effects of multipath propagation as traditional SISO (Single-Input Single-Output) communication systems [1], [8]

Multi-user multi-antenna transmission architecture with channel estimators cascaded at the receiver side is proposed so that each user can feedback channel state information (CSI) for the further process of antenna resource allocation [2][3].

In MIMO, "multiple in" means a WLAN device simultaneously sends two or more radio signals into multiple transmitting antennas. "Multiple out" refers to two or more radio signals coming from multiple receiving antennas. These views of "in" and "out" may seem reversed; but MIMO terminology focuses on the system interface with antennas rather than the air interface. Whatever be the terminology, the MIMO's basic advantage seems simple, i.e. multiple antennas receive more signal and transmit more signal [1],[5],[8]. Maximal receive combining takes the signals from multiple antennas/receivers and combines them in a way that significantly boosts signal strength[6]. This technique is fully compatible with standard 802.11a/b/g. It significantly improves overall gain, especially in multipath environments.

In multipath environments, signals pass through and reflect from various objects so that different signal reaches the two receiving antennas. Some frequencies tend to be attenuated at one antenna but not the other, which is shown by channel measurements in a multipath environment [5],[7]. The capacity of the phased array system grows logarithmically with increasing antenna array size, whereas the capacity of the MIMO system grows linearly[10],[15].

II. MIMO SYSTEM

A. MIMO wireless system

MIMO wireless system consists of two antennas $N \times M$. N antennas transmit the data whereas M antennas are to receive the data. MIMO system is different from other phased array systems where a single information stream, say $x(t)$, is transmitted on all transmitters and then received at the receiver antennas. It can transmit different information streams $x(t)$, $y(t)$, $z(t)$, on each transmit antenna. These are independent information streams being sent simultaneously and in the same frequency band. The received signals $r_1(t)$, $r_2(t)$, $r_3(t)$ at each of the three received antennas are a linear combination of $x(t)$, $y(t)$, $z(t)$ [6],[8]. The coefficients $\{a_{ij}\}$ represent the channel weights corresponding to the attenuation seen between each transmit-receive antenna pair. The affect is that we have a system of three equations and three unknowns as shown below.

$$R = A [x \ y \ z]$$

The matrix, A , of channel coefficients $\{a_{ij}\}$ must be invertible for MIMO systems to live up to their promise. It has been proven that the likelihood for A to be invertible increases as the number of multipaths and reflections in the vicinity of the transmitter or receiver increases. The impact of this is that in a Rayleigh fading environment with spatial independence, there are essentially NM levels of diversity available and there are $\min(N,M)$ independent parallel channels that can be established. Increases in the diversity order results in significant reductions in the total transmit power for the same level of performance[15]. On the other hand, an increase in the number of parallel channels translates into an increase in the achievable data rate within the same bandwidth.

B. MIMO Techniques

There are four unique multi-antenna MIMO techniques available to the system designer namely : spatial

multiplexing (SM-MIMO), space-time coding (STC-MIMO), diversity systems (DIV-MIMO), smart antenna (SA-MIMO): In spatial multiplexing, a high rate signal is split into multiple lower rate streams and each stream is transmitted from a different transmit antenna in the same frequency channel. If these signals arrive at the receiver antenna array with sufficiently different spatial signatures, the receiver can separate these streams, creating parallel channels free. Spatial multiplexing is a very powerful technique for increasing channel capacity at higher Signal to Noise Ratio (SNR)[6]. The maximum number of spatial streams is limited by the lesser in the number of antennas at the transmitter or receiver. Spatial multiplexing can be used with or without transmit channel knowledge. Spatial multiplexing MIMO schemes have been suggested to solve any and all wireless communication issues. Spatial multiplexing maximizes the link capacity, for spatial multiplexing the number of receive antennas must be greater than or equal to the number of transmit antennas [8]. It makes the receivers very complex, and therefore it is typically combined with orthogonal frequency-division multiplexing (OFDM) [1], [4]. The IEEE 802.16e standard incorporates MIMO-OFDMA. The IEEE 802.11n standard which is expected to be finalized soon, recommends MIMO-OFDM. Compared to spatial multiplexing systems, space-time code STC-MIMO systems provide robustness of communications without providing significant throughput gains against spatial multiplexing systems [6], [13].

Moreover, to support fully the cellular environments MIMO research consortiums including IST-MASCOT, proposed to develop advanced MIMO communication techniques such as cross-layer MIMO, multi-user MIMO and ad-hoc MIMO. Cross-layer MIMO enhances the performance of MIMO links by solving cross-layer problems occurred when the MIMO configuration is employed in the system. A Cross-layer technique has been enhancing the performance of SISO links as well [7]. Examples of cross-layer techniques are Joint source-channel coding, Link adaptation, or adaptive modulation and coding (AMC), Hybrid ARQ (HARQ) and user scheduling. Multi-user MIMO can exploit multiple user interference powers as a spatial resource at the cost of advanced transmit processing while conventional or single-user MIMO uses only the multiple antenna dimension[4]. Examples of advanced transmit processing for multi-user MIMO are interference aware precoding and SDMA-based user scheduling.

Ad-hoc MIMO is a useful technique for future cellular networks which considers wireless mesh networking or wireless ad-hoc networking. To optimize the capacity of ad-hoc channels, MIMO concept and techniques can be applied to multiple links between transmit and receive node clusters. Unlike multiple antennas at the single-user MIMO transceiver, multiple nodes are located in a distributed manner. So, to achieve the capacity of this network, techniques to manage distributed radio resources are essential like the node cooperation and dirty paper coding (DPC) [8].

III. MIMO Channel

New transmit strategies are derived and compared to existing transmit strategies, such as beamforming and space-time block coding (STBC). Rayleigh fading multiple input multiple output (MIMO) channels are studied using an eigenvalue analysis and exact expressions for the bit error rates and outage capacities for beamforming and STBC is found[6]. In general are MIMO fading channels correlated and there exists a mutual coupling between antenna elements. These findings are supported by indoor MIMO measurements. It is found that the mutual coupling can, in some scenarios, increase the outage capacity[9]. The effects of nonlinear transmit amplifiers in array antennas are also analyzed, and it is shown that an array reduces the effective intermodulation distortion (IMD) transmitted by the array antenna by a spatial filtering of the IMD. The use of a low cost antenna with switchable directional properties, the switched parasitic antenna, is studied in a MIMO context and compared to array techniques. It is found that it has comparable performance, at a fraction of the cost for an array antenna. In recent years, deploying multiple antennas at both transmitter and receiver has appeared as a very promising technology[8]. By exploiting the spatial domain, multiple-input multiple-output (MIMO) systems can support extremely high data rates as long as the environments can provide sufficiently rich scattering. To design high performance MIMO wireless systems and predict system performance under various circumstances, it is of great interest to have accurate MIMO wireless channel models for different scenarios.

A. Space-time block code

Space-time block coding is a technique used to transmit multiple copies of a data stream across a number of antennas and to exploit the various received versions of the data to improve the reliability of data-transfer Alamouti invented the simplest of all the STBCs. It is readily apparent that this is a rate-1 code. It takes two time-slots to transmit two symbols. Using the optimal decoding scheme discussed below, the bit-error rate (BER) of this STBC is equivalent to 2nR-branch maximal ratio combining (MRC)[13]. This is a very special STBC. It is the only orthogonal STBC that achieves rate-1. That is to say that it is the only STBC that can achieve its full diversity gain without needing to sacrifice its data rate[13]. Strictly, this is only true for complex modulation symbols. Since almost all constellation diagrams rely on complex numbers however, this property usually gives Alamouti's code a significant advantage over the higher-order STBCs even though they achieve a better error-rate performance [14].

Tarokh et al, discovered a set of STBCs that are particularly straightforward, and coined the scheme's name. They also proved that no code for more than 2 transmit antennas could achieve full-rate. They also demonstrated the simple, linear decoding scheme that goes with their codes under perfect channel state information assumption [16]. One particularly attractive feature of orthogonal STBCs is that maximum

likelihood decoding can be achieved at the receiver with only linear processing.

We have calculate the error probability achieved by the MRC, showing it to be much smaller than the one corresponding to the SISO channel, in which no spatial diversity exists. Next, we consider the multiple-input single-output (MISO, multiple transmit antennas, single receive antenna) channel, and we present some mechanisms that exploit the transmit diversity offered by this channel. Specifically, Alamouti's scheme is analyzed. Bringing together transmit and receive diversity, the MIMO channel is introduced. The Alamouti-based scheme are shown to achieve full diversity, i.e., they take full advantage of both transmit and receive diversity provided by the MIMO channel.

IVSCHEDULING ALGORITHM

Multiuser scheduling is the problem of allocating resources (such as power or bandwidth) in order to perform desirably with respect to criteria such as throughput or delay. Most previous studies limit their scope to time-sharing schedules. Transmitting to the user with the best reception is sum-rate optimal (achieves maximum throughput) for a single-antenna broadcast channel under infinite backlogs and symmetric channels. However, in a multiple-antenna broadcast channel time-division is sub-optimal.

Schedules most commonly ignore queuing and randomness in packet arrivals and hence cannot offer stability guarantees. This is true in some scheduling algorithms that aim to satisfy fairness criteria, such as proportional-fair scheduling:

B. Multiuser scheduling

Multiuser scheduling is the problem of allocating resources (such as power or bandwidth) in order to perform desirably with respect to criteria such as throughput or delay. This problem has attracted great interest in the recent years. Most previous studies limit their scope to time-sharing schedules, i.e. those where only a single user's data is transmitted at any time. The computational complexity of broadcast coding, together with the fact that the optimal coding for the MIMO Broadcast channel was not known until recently, has made time-sharing attractive. In fact, transmitting to the user with the best reception is sum-rate optimal (achieves maximum throughput) for a single-antenna broadcast channel under infinite backlogs and symmetric channels. However, in a multiple-antenna broadcast channel time-division is sub-optimal. Schedules proposed in previous literature also most commonly ignore queuing and randomness in packet arrivals and hence cannot offer stability guarantees. This is true in some scheduling algorithms that aim to satisfy a fairness criteria, such as proportional-fair scheduling.

A guiding work for incorporating randomness and stability issues has been, where the network capacity region is defined as the region of stabilizable input data rates, and it is shown that this region is achieved by a maximum-weight matching (weights being related to queue sizes). Building on those definitions, [11] considers a broadcast scenario under time

division and demonstrates a schedule that achieves the network capacity region. Along similar lines, [12] shows that a throughput-optimal policy is a maximum-weight matching in the form of $\max \sum q_i r_i$ where q_i 's are the queue states of users, and the rates r_i are left implicit. Also in the downlink scenario, compares several heuristic scheduling policies such as beamforming to the user with the shortest remaining job versus multiplexing several users. To our knowledge, the maximum-weight matching scheduling policy has first been combined with channel coding and power control explicitly in [14], for the multi-access channel.

C. Scheduling Policies

The Most of the current researches focus on the fairness among users. Nevertheless, it has been found however that a dilemma exists between fairness and system capacity arrangement. The goal of fairness scheduling is to deliver the equal information bits to users, while that of capacity scheduling is to maximize the utilization of wireless channels. The best way to achieve highest overall throughput of the system is to assign higher data rate for those subchannels in good condition, and to assign lower data rate for those poor subchannels. Unfortunately, under multi-user system architecture, each subchannel stands for each user so favoring particular subchannel leads to unfairness issue.

D. Antenna Scheduling and selection

For the wireline communications, several scheduling techniques such as weighted fair queuing and packetized general processor sharing have been proposed to furnish fair channel access among contending hosts. However, an attempt to apply these wireline scheduling algorithms to wireless systems is inappropriate because wireless communication system presents many new challenges such as radio channel impairments. Therefore, late researches investigate some resources such as code, power, and bandwidth to exploit more efficient transmission under wireless MIMO environment [11], [12], [13]. We explore an antenna allocation scheme with dynamic allotment of multiple antennas for each real-time user to satisfy their QoS requirements. Although fairness is an important criterion in judging the design of a scheduling algorithm. Overemphasizing it is not good in reality because "fairness" does not equal to user's satisfaction. Hence we propose a different algorithm which targets to satisfy user's QoS by allocating the number of transmit antennas. In this algorithm, we have to calculate how many antennas a user should use in order to satisfy user's time-varying data rate requests. Since we assume that the SNR and spatial correlation are known at the transmitter and the receive antenna amounts are naturally known. So we can compute the channel capacity as the function of the number of transmit antenna. Which antenna should be added or taken off as the next step would be dependent upon how many antennas are to be used.

VI SIMULATIONS RESULT

To assess the relative performance of the MIMO system, we consider as metrics the latency, fairness and average rate In Fig. 1

we plot the Channel capacity (in total number of bits per second per hertz (b/s/Hz)) and the SNR of the downlink channel as a function of the number of transmitting (nt)and receiving antennas(nr), respectively. It should be noted that the use of multiple antennas has significant impact on the Channel Capacity. We have calculate the error probability achieved by the MRC, showing it to be much smaller than the one corresponding to the SISO channel, in which no spatial diversity exists. Next, we consider the multiple-input single-output (MISO, multiple transmit antennas, single receive antenna) channel, and we present some mechanisms that exploit the transmit diversity offered by this channel. Specifically, Alamouti's schemes are analyzed. In Fig. 2 and 3, we plot SNR vs BER for 2 Transmitter &1Reciver and 2 Transmitter &2Reciver,it is evident from plot that BER is minimum in the case of MIMO system. The Alamouti-based scheme are shown to achieve full diversity, i.e., they take full advantage of both transmit and receive diversity provided by the MIMO channel.

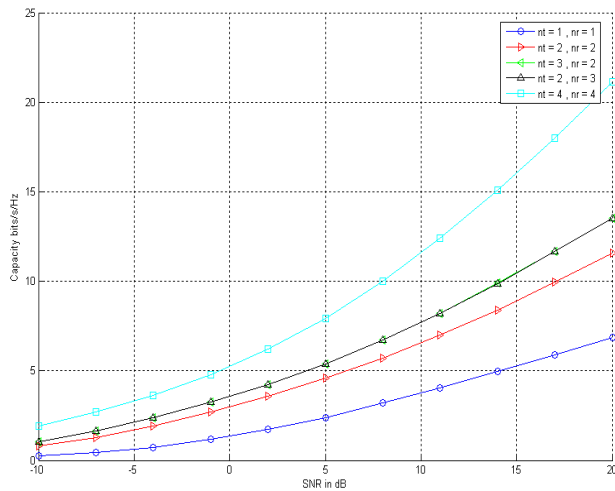


Figure 1. Channel Capacity versus SNR

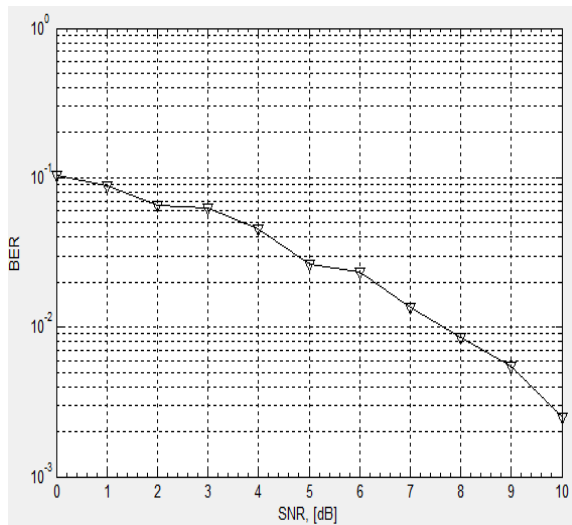


Figure 2. BER vs SNR (2 Transmitter &1Reciver)

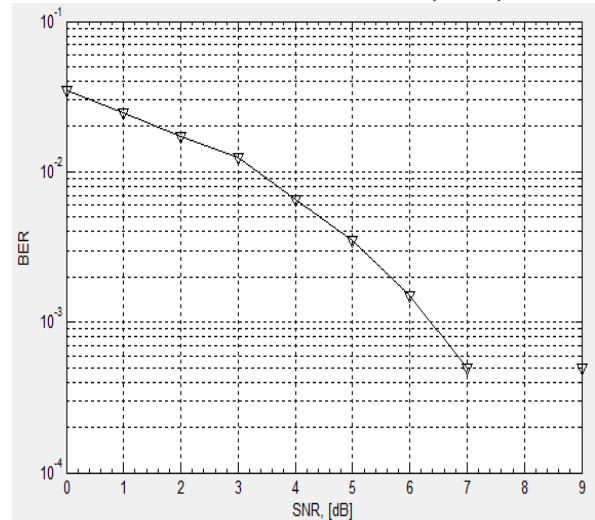


Figure 3. BER Vs SNR(2 Transmitter &2Reciver)

The scheduling performance of our algorithm under two different types of traffic mode are implemented: one is voice or web-browsing in that bursts of data rate happen in some time intervals, sometimes it occurs silently also. The other one is for data transfer and streaming data. The requirement of data is self-similar and constantly high or low with only few fluctuations. The former is modeled by Pareto distribution while the later one is modeled by Weibull distribution. In the following simulations, channel matrix change every 10 time index with a total 12 transmit antennas for 3 users, and the algorithm trigger threshold at 1.5 bits/Hz/sec. The data streaming traffic mode is modeled by Weibull distribution with given pdf,

$$f(x) = aBx^{B-1} e^{-axB}$$

Where a is the scale parameter and B is the shape parameter. The pdf distribution is shown in Figure 6.3. Data rate request and indemnity curves for Weibull Distribution are shown in Figure 4, Figure 5, and Figure 6 respectively.

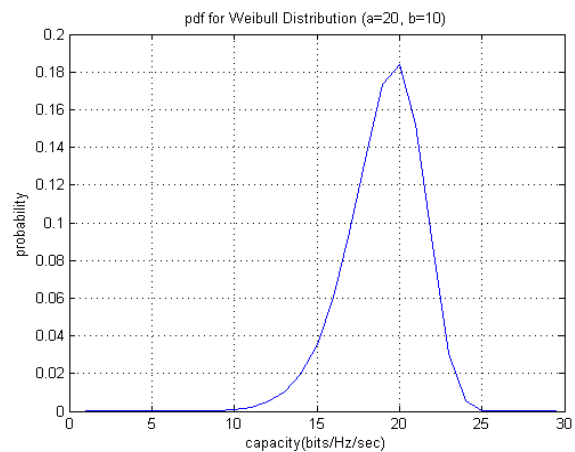


Figure 4: Weibull Distribution

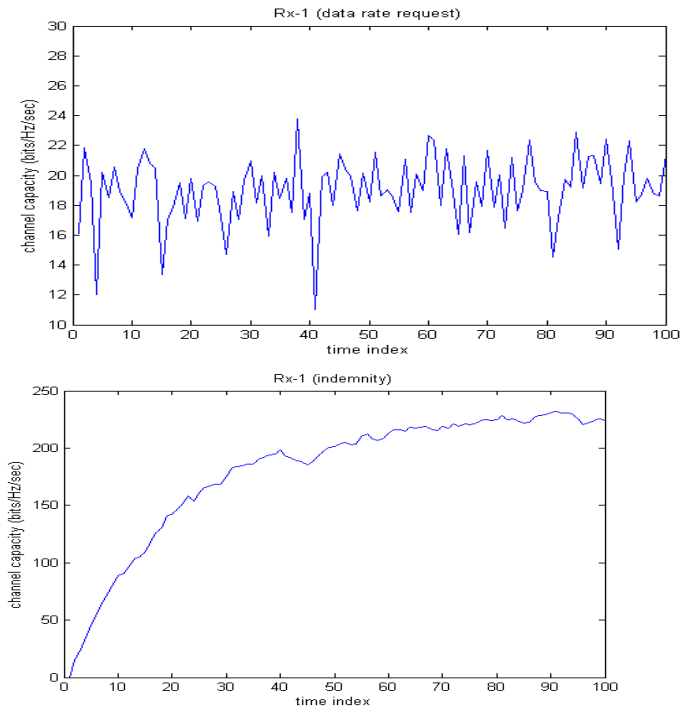


Figure 5: Data Rate Request and Indemnity Curves for Weibull Distribution ($a=20, b=10$) with $R_x=1$

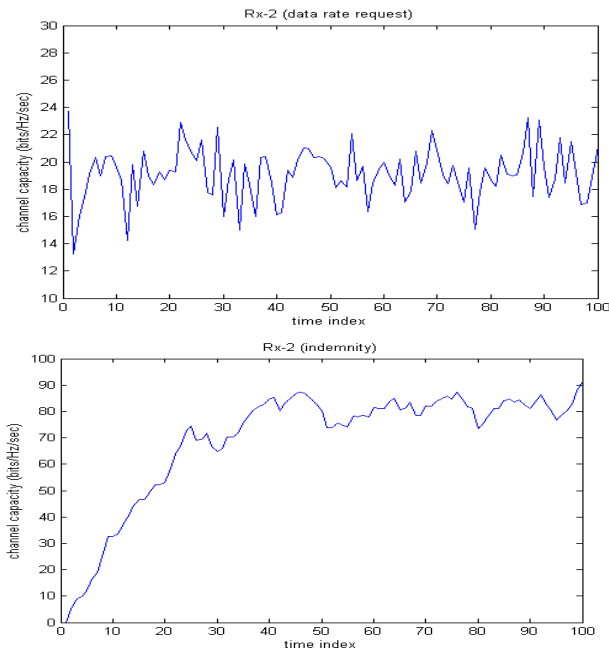


Figure 6: Data Rate Request and Indemnity Curves for Weibull Distribution ($a=20, b=10$) with $R_x=2$

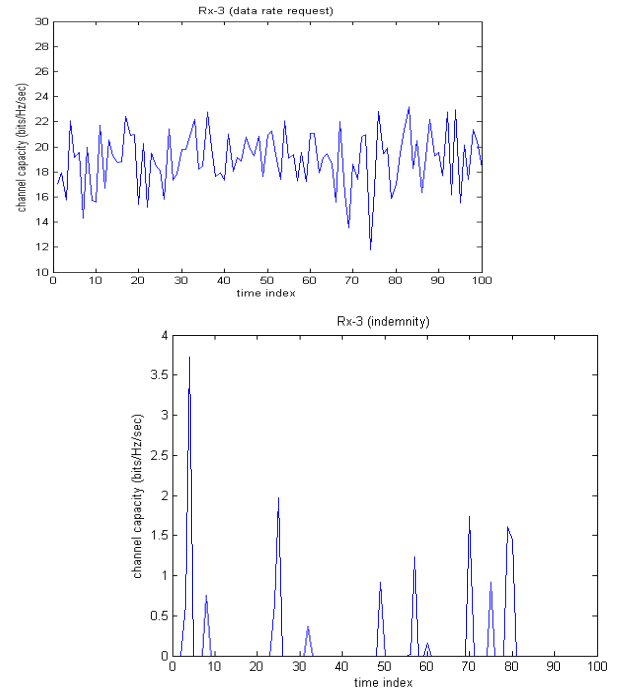


Figure 7: Data Rate Request and Indemnity Curves for Weibull Distribution ($a=20, b=10$) with $R_x=3$

Pareto distribution

The former described traffic flow is modeled by Pareto distribution with given pdf as

$$f(x) = B a^B / x^{B+1}$$

Where a is the scale parameter and B is the shape parameter.

The pdf distribution is shown in Figure 8. Data rate request and indemnity curves for Pareto Distribution are shown in Figure 9, Figure 10 and Figure 11 respectively.

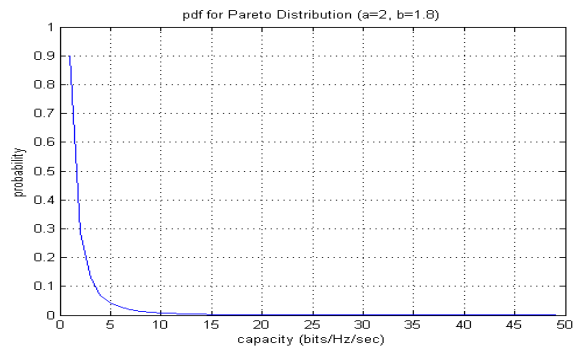


Figure 8: Pareto Distribution

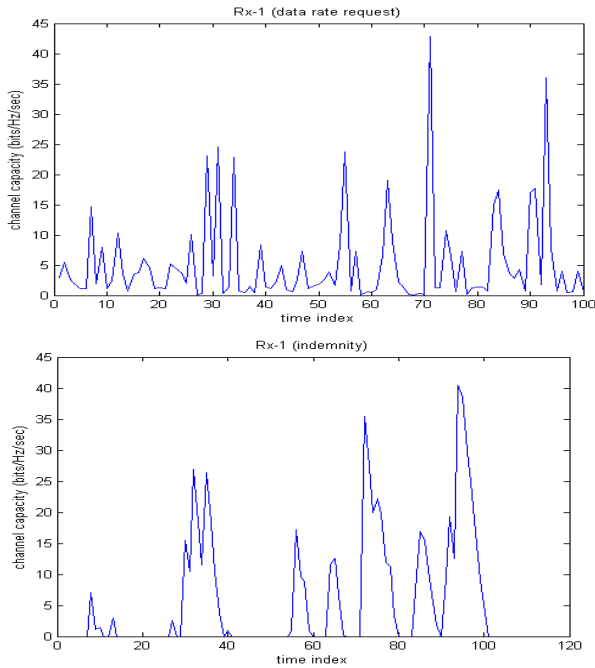


Figure 9: Data Rate Request and Indemnity Curves for Pareto Distribution ($a=1.8, b=2$) with $R_x=1$

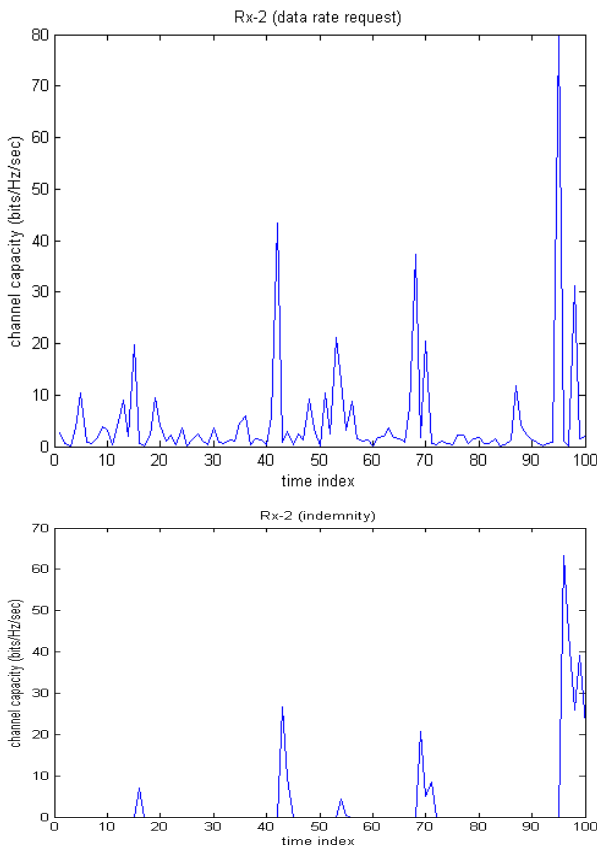


Figure 10: Data Rate Request and Indemnity Curves for Pareto Distribution ($a=1.8, b=2$) with $R_x=2$

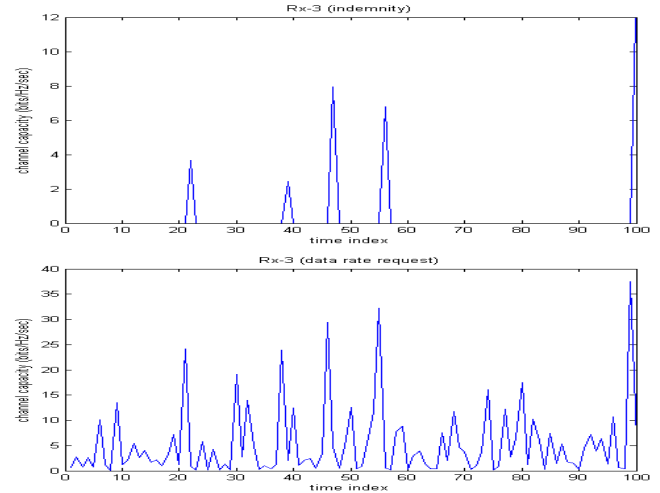


Figure 11: Data Rate Request and Indemnity Curves for Pareto Distribution ($a=1.8, b=2$) with $R_x=3$

Weibull distribution traffic mode is for high-data rate transmission (with average throughput request about 20 bits/HZ/sec); so using single receive antenna is not enough for handling the constantly high data rate and in the end it has compensation diverge, i.e. system crashes depicted in Figure 5. Changing forgetting factor smaller could somehow relieve the traffic pressure shown in Figure 11, but it doesn't solve the problem from the bottom line. Four receive antennas is suggested at least for such high-data rate system requirement. Though applying smaller forgetting factors can alleviate the divergence of compensation, the system turns out to be sensitive to the sudden change of 'data rate request' particularly for the case of Weibull distribution. The indemnity can avoid numbers of antennas being taken off when the data rate request abruptly drops; for the sake of this, the system doesn't have to increase the number of transmit antennas when the data rate request goes back to normal. What's more, excessive small forgetting factor equals to no compensation. So by choosing an appropriate forgetting factor more users can be accommodated by the system. Either the higher signal-to-noise ratio or the less correlation, offer a better environment for transmission and exploration of more capacities. In the time domain analysis, we also evaluate how many transmit antennas are needed to reach certain service quality (to guarantee the average indemnity under certain level).

VII. CONCLUSION

In this paper, broader scope of MIMO channel modeling methods was presented. Through simulation study, we showed that, how many transmit antennas are needed to reach certain service quality and Either the higher signal-to-noise ratio or the less correlation, offer a better environment for transmission and exploration of more capacities. We briefly went through diverse scheduling policies and proposed a novel, different optimizing target of antenna selection and scheduling. Simultaneously, results of antenna scheduling

algorithm under random traffic mode-Weibull and Pareto-are discussed. It is possible that the future high data-rate-demand communication system would require more transmit antennas to overcome the bottleneck of limited capacity. Our algorithm requires as many transmit antennas as that of receive antennas approximately and it would be very helpful to enhance the overall usage of channel resource for high-speed wireless network physical layer design.

REFERENCES

- [1] H. Yang, "A road to future broadband wireless access: MIMO-OFDM based air interface," IEEE Comm. Mag., vol. 43, no. 1, pp. 53–60, Jan.2005.
- [2] Ying Jun Zhang and Khaled Ben Letaief, "Adaptive Resource Allocation for Multiaccess MIMO/OFDM Systems With Matched Filtering", IEEE Transactions on Communication, VOL. 53, NO. 11, pp 1810-1816, Nov 2005.
- [3] Alkan Soysal and Sennur Ulukus, "Joint Channel Estimation and Resource Allocation for MIMO Systems. IEEE Transactions on Wireless Communication, vol. 9, no. 2, pp. 632-640 Feb 2010.
- [4] Y.-J. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," IEEE Trans. Wireless Commun., vol. 3, no. 5, pp. 1566–1575, Sep. 2004.
- [5] Atheros Communications, Inc., "Getting the Most out of MIMO: Boosting Wireless LAN Performance with Full Compatibility.", pp 3-9, June 2005.
- [6] Shahram Shahbazpanahi, et al "Minimum Variance Linear Receivers for Multiaccess MIMO Wireless Systems with Space-Time Block Coding" IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 52, NO. 12, DECEMBER 2004.
- [7] H. Jiang, W. Zhuang, and X. Shen, "Cross-layer design for resource allocation in 3G wireless networks and beyond," IEEE Comm. Mag., vol. 43, no. 12, pp. 120-126, Dec. 2005.
- [8] Shailendra Mishra and D.S.chauhan, "Link Level Performance of Multiple input Multiple Output (MIMO) System " International Journal of Control and Automation, Vol.2, No.3, pp 29-41, September 2009.
- [9] Chen-Nee Chuah, David N. C. Tse, Joseph M. Kahn and Reinaldo A. Valenzuela, "Capacity Scaling in MIMO Wireless Systems under Correlated Fading." IEEE Transactions on information theory, vol. 48, no. 3, March 2002
- [10] I. G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," IEEE Trans. Inform. Theory, vol. 49, no. 7, pp. 1691–1706, July 2003.
- [11] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," Euro. Trans.Telecommun., vol. 10, no. 6, pp. 585–96, Nov. 1999
- [12] G. J. Fochini, "Layered Space-time Architecture for Wireless Communication in Fading Environment when Using Multi-Element Antennas," Bell Labs. Tech.J., vol. 1, pp. 41–59, Nov. 1996
- [13] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-Time Block Codes from Orthogonal Designs," IEEE Trans. Info. Theory, vol. 45, no. 5, pp. 1456–67, July 1999
- [14] S. M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," IEEE Journal on Selected Areas in Communications, vol. 16, pp 1451-1458, October 1998
- [15] Babak Daneshrad, "MIMO: The next revolution in wireless data communications" DefenseElectronics, RF Design www.rfdesign.com
- [16] Tarokh Vahid, Hamid Jafarkhani, and A. Robert Calderbank, "Space-Time Block Coding for Wireless Communications: Performance Results" IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 17, NO. 3, pp-451-461 MARCH 1999

AUTHORS PROFILE



Shailendra Mishra

Received Ph.D degree and Master of Engineering Degree (ME) in Computer Science & Engineering (Specialization: Software Engineering) from Motilal Nehru National Institute of Technology (MNNIT), Allahabad India(2000) and B.Sc degree from University of Allahabad, India in 1990 respectively. From August 1994 to Feb 2001, he was associated with the Department of Computer Science and Electronics at ADC, University of Allahabad, India as a faculty and coordinator Computer Science & Electronics Department. From Feb 2001 to Feb 2002 he has been with RG Engineering College, Meerut affiliated to UP Technical University, Lucknow, India as Assistant Professor, Department of Computer Science and Engineering. From Feb 2002 to Aug.2009 he had been with Dehradun Institute of Technology, Dehradun as Professor, Deptt. of Computer Science & Engineering. Presently he is Professor & Head, Department of Computer Science & Engineering Kumaon Govt. Engineering College, Dwarahat, Uttarakhand India. His recent research has been in the field of Mobile Computing & Communication and Advance Network Architecture. He has also been conducting research on Communication System & Computer Networks with Performance evaluation and design of Multiple Access Protocol for Mobile Communication Network. He handled many research projects during the last 5 years; Power control and resource management for WCDMA System funded and sponsored by AICTE New Delhi, Code and Time complexity for WCDMA System, OCQPSK spreading techniques for third generation communication system, "IT mediated education and dissemination of health information via Training & e-Learning Platform" sponsored and funded by Oil Natural Gas Commission (ONGC), New Delhi, India (November 2006), "IT based Training and E-Learning Platform", sponsored and funded by UCOST, Department of Science and Technology, Govt. of Uttarakhand, India (December 2006) etc. He authored two books in the area of Computer Network and Security and published 30 research papers in International journals and International conference proceedings.



Prof. Durg Singh Chauhan

He did his B.Sc Engg. (1972) in electrical engineering at I.T. B.H.U., M.E. (1978) at R.E.C. Tiruchirapalli (Madras University), PH.D. (1986) at IIT, Delhi and his post doctoral work at Goddard space Flight Centre, Greenbelt Maryland, USA (1988-91). His brilliant career brought him to teaching profession at Banaras Hindu University where he was Lecturer, Reader and then has been Professor till today. He was director KNIT Sultanpur in 1999-2000 and founder vice Chancellor of U.P.Tech. University (2000-2003-2006). Later on, he served as Vice-Chancellor of Lovely Profession University (2006-07) and Jaypee University of Information Technology (2007-2009) currently he has been serving as Vice-Chancellor of Uttarakhand Technical University for (2009-12) Tenure. He was member, NBA-executive AICTE, (2001-04)-NABL-DST executive (2002-05) and member, National expert Committee for IIT-NIT research grants. Currently He is Member, University Grant Commission (2006-09). He is presently nominated by UGC as chairman of Advisory committees of three medical universities. Dr Chauhan got best Engineer honour of institution of Engineer in 2001 at Lucknow. He supervised 12 Ph.D., one D.Sc and currently guiding half dozen research scholars. He had authored two books and published and presented 85 research papers in international journals and international

international conferences and wrote more than 20 articles on various topics in national magazines. He is Fellow of institution of Engineers and IEEE.

Email Authorship Identification Using Radial Basis Function

A.Pandian
Asst.Professor (Senior Grade)
Department of MCA
SRM University, Chennai, India
pandiana@ktr.srmuniv.ac.in

Dr. Md. Abdul Karim Sadiq
Ministry of Higher Education
College of Applied Sciences, Sohar,
Sultanate of Oman
abdulkarim.soh@cas.edu.om

Abstract - Email authorship identification helps tracking fraudulent emails. This research proposes extraction on unique words from the emails. These unique words will be used as representative features to train Radial Basis function (RBF). Final weights are obtained and subsequently used for testing. The percentage of identification of email authorship depends upon number of RBF centers and the type of functional words used for training RBF. One hundred fifty authors with one hundred files from the sent folder of Enron database are considered. A total of 300 unique words of number of characters in each word ranging from 3 to 7 are considered. Training and Testing RBF are done by taking different length of words. The percentage of authorship identification ranges from 95% to 97%. Simulation shows the effectiveness of the proposed RBF network for email authorship identification.

Keywords: email authorship identification; word frequency; radial basis function;

I. INTRODUCTION

The principal objectives of author identification are to classify [Moshe 2002] the emails belonging to an author. This approach is used in forensic for author identification in malicious emails. Some of the commercial softwares like copycatch gold, jvocalize, signature stylometric system, textaz, Antconc, yoshikoder, lexico3, T-lab, wordsmithtools etc. use statistical methods to identify an author.. These softwares uses parameters such as total number of different words, number of content words used in the list, total number of words in the text / vocabulary items used, vocabulary richness, mean sentence length, mean paragraph length, mean of 2-3 letter words, mean of voxel starting words, cumulative summation method, bigrams and many more. The users who intend to utilize the software for their email author identification need to choose the type of statistical analysis options that best identify author for an email and obtain the characteristics that remains constant for large number of emails written

by the author. Each author follows style, which is called functional words. By using these functional words and their frequencies, identification of the author is easy [David 2005].

Authorship identification is important as the number of documents in internet is increasing. The researchers are focused on different properties of texts. There are two different properties of the texts that are used in classification: the content of the text and the style of the author. Stylometry [Goodman 2007] the statistical analysis of literary style - complements traditional literary scholarship since it offers a means of capturing the often elusive character of an author's style [Zheng 2006] by quantifying some of its features. Most stylometry [Pavelec 2007 and Diederich 2008] studies employ items of language and most of these are lexically based.

The usefulness of function words in Authorship attribution [Diederich 2003] is examined. Experiments were conducted with support vector machine classifiers in twenty novels and-success rates above 90% were obtained. The use of functional words is a valid and good approach in Authorship attribution [Koppel 2006].

Stamatatos 2001 has measured a success rate of 65% and 72% in their study for authorship recognition, which is an implementation of multiple regression and discriminant analysis. Joachim Diederich 2003 and his collaborators conducted experiments with support vector classifiers and detected author with 60-80% success rates with different parameters.

The effect of word sequences in authorship [Abbasi 2005] attribution has been studied. The researchers aimed to consider both stylistic and topic features of texts. In this work the documents are identified by the set of word sequences that combine functional and content words. The experiments are done on a dataset consisting of poems using naïve

Bayes classifier [Peng 2004]; the researchers claim that they achieved good results.

II. MATERIALS AND METHODS

2.1 Materials

Words of working type, action oriented, different categories of prepositions, pronouns, adjectives, adverbs, conjunctions and interjections are given in Table 1 to Table 3. These words are used as filtering and as templates. When an email is analyzed for uniqueness, the extracted features are based on list of words presented in the tables. Hence, unnecessary words are eliminated and the number of unique words that represent an email is minimum.

TABLE 1 SAMPLE WORDS USED FOR FILTERING

Work (70)	Action (524)	Preposition 1 (94)	Preposition_2 (30)
analyze	Accelerate	Aboard	according to
annotate	Accommodate	About	ahead of
ascertain	Accomplish	Above	as of
attend	Accumulate	Absent	as per
audit	Achieve	Across	as regards
build	Acquire	After	aside from
calculate	Act	Against	because of
consider	Activate	Along	close to
construct	Adapt	Alongside	due to
control	Add	Amid	except for

TABLE 2 SAMPLE WORDS USED FOR FILTERING

Preposition 3 (16)	Preposition 4 (9)	Pronoun (77)	Adjectives (395)
as far as	apart from	All	early
as well as	but	Another	abundant
by means of	except	Any	adorable
in accordance with	plus	anybody	adventurous
in addition to	save	Anyone	aggressive
in case of	concerning	anything	agreeable
in front of	considering	Both	alert
in lieu of	regarding	Each	alive
in place of	worth	each other	amused
in point of		Either	ancient

TABLE 3 SAMPLE WORDS USED FOR FILTERING

Adverbs (331)	Conjunctions (25)	Interjections (77)
Abnormally	And	Absolutely
absentmindedly	But	Achoo
Accidentally	For	Ack
Acidly	Nor	Agreed
Actually	Or	Aha
Adventurously	So	Ahem
Afterwards	Yet	Ahh
Almost	after	Ahoy
Always	although	Alack
Angrily	as	Alas

Work words: To avoid misinterpretation, work words will analyze how an author writes his email

and what clarity he has in the mail. The number of work words will indicate performance task requirements in a neat, unambiguous manner by using the work words that translate exactly what an author has in his mind. Action words: It indicates some actions during an expressing in the email. Preposition, adjectives, adverbs, conjunctions and interjections have their standard meanings.

The total number of words used as basic dictionary is 1648 (work + action + prepositions + adjectives + adverbs + conjunctions + Interjections). The numbers mentioned in the paranthesis are the total in each category whereas, only few words are shown in the tables for understanding.

A schematic diagram for implementation of the proposed work is presented din Figure 1.

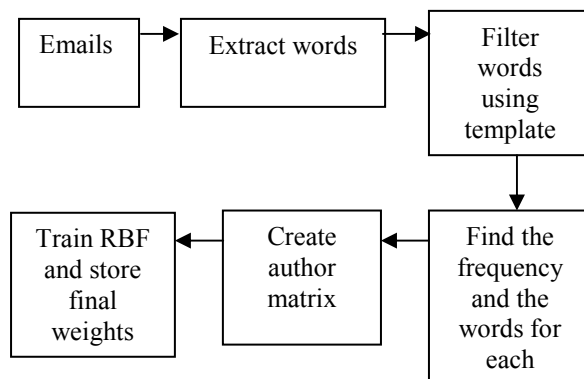


Fig.1 (a) Training the system

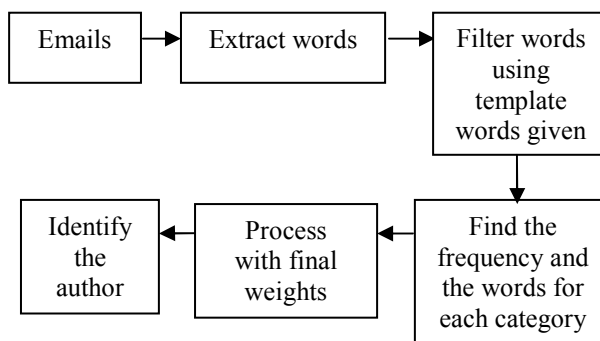


Fig.1 (b) Testing the system

Email: The email received in the system

Extract words: all the words in the email are arranged.

Filter words: The words given in Table 1-3 are searched in the extracted words. Subsequently, the word frequencies are found.

Author matrix: A matrix with column as authors and vertical rows with word frequencies.

Training patterns: The columns of the matrix are used as training patterns and labeling are introduced.

2.2 Methods

2.2.1 Radial Basis Function

The concept of distance measure is used to associate the input and output pattern values. RBFs are capable of producing approximations to an unknown function 'f' from a set of input data abscissa. The approximation is produced by passing an input point through a set of basis functions, each of which contains one of the RBF centers.

An exponential function is used as an activation function for the input data. Distance between Input data and set of centers chosen from the Input data are found and passed through an exponential activation function. A bias value of f is used along with the data. These data are further processed to get a set of final weights between radial basis function and the target value.

The topology of RBF network is 12 nodes in Input layer, 10 nodes in hidden layer and 1 node in the output layer. The difference in input data and a center is passed through $\exp(-x)$ and is called RBF. A rectangular matrix is further obtained for which inverse is found. The resultant value is processed with the entire inputs and target values to obtain final weights.

Details of the Figure 2 is given below:

Read input pattern: The columns of the author matrix are used as training patterns. The number of patterns is equal to number of authors.

Create center: One hundred training patterns are used as centers.

Create RBF: Calculate distance between training patterns and one hundred centers. The resultant values are passed through activation function, $\exp(-x)$ to produce outputs of RBF nodes in the hidden layer of the network.

The number of training patterns and the number of centers will produce a rectangular matrix. This is converted into square matrix and inverse of the same is found and processed with labeling to get final weights.

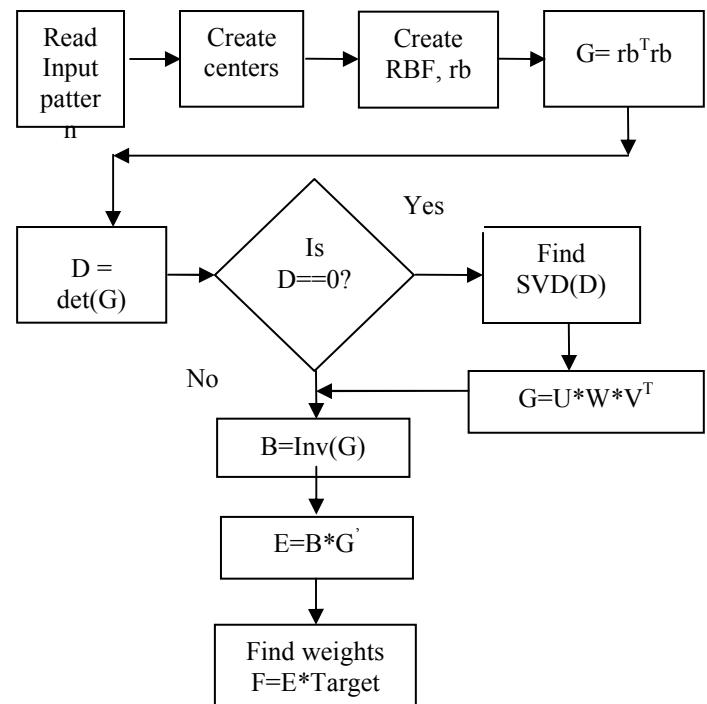


Fig 2 Radial basis function flow chart

III. EXPERIMENTAL PROCEDURE

Enron email dataset has been used for evaluating the efficiency of RBF in email authorship identification. This email dataset was made public by the Federal Energy Regulatory Commission during its investigation. It contains all kind of emails, personal and official. William Cohen from CMU has put up the dataset on the web for researchers. This contains around 5,17,431 emails from 151 users. Each mail in the folders contains the senders and the receiver email addresses, date and time, subject, body, text and some other email specific technical details. It is available in the form of MySQL database with a size of 400MB. The Enron database contains four tables. The first table contains information of each of the 151 employees. The second table contains the information of the email message, the sender, subject, text and other information. The third table contains the recipient's information. The fourth table contains information about either as a forward or reply. Table 4 presents names of few folders under each author. Only 146 authors have been considered for study.

TABLE 4 DETAILS OF ENRON FOLDER

Person	Sent_mail	All documents	contacts	Deleted_items	Discussion_threads	inbox	Notes inbox	sent	sent_items
allen-p	602	628	2	361	412	66	48	562	345
arnold-j	814	1047	X	723	401	142	84	816	723
arora-h	X	65	X	197	57	79	X	9	68
badeer-r	52	299	2	13	277	3	115	X	7
bailey-s	X	16	X	434	X	4	10	X	14
bass-e	1409	2037	X	415	1386	310	601	1363	258
Baughman-d	X	389	4	431	384	383	X		96
beck-s	1093	3137	7	309	2630	751	190	1099	482
benson-r	X	84	X	203	77	274	75	7	9
blair-l	39	2	X	662	X	291	X	X	929

X represents no information

There are 15 unique words that are identified in all the emails under consideration by using the filtering words given in Table 1-3.. The list of unique words is presented in Table 5.

TABLE 5 UNIQUE WORDS

our	when
out	which
plan	with
please	you
that	your
to	yours
we	zip
what	

IV. IMPLEMENTATION

Characterization and feature extraction for training radial basis function (RBF) are based on vowels at the beginning of words and some of the grammatical rules present in the emails of an author. Figure 3 presents authors in x-axis and number of words with vowels at the beginning of words in the y-axis. Each stem is the average number of words with vowels at the beginning considering all the emails of an author. Figure 4 presents the number of work words. Figure 5 presents the number of action words. Figure 6 to figure 9 presents number of prepositions.

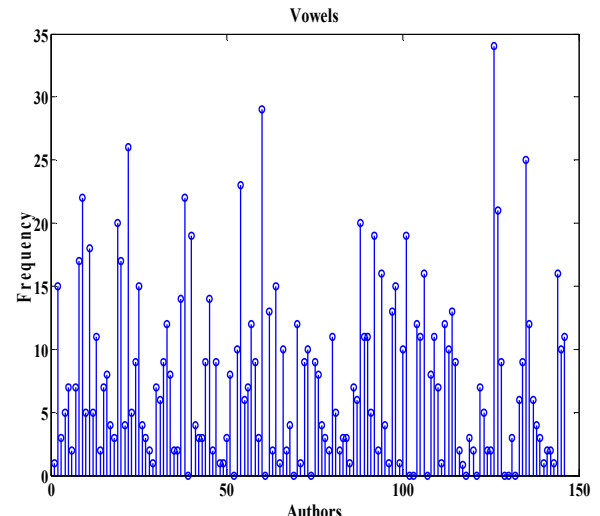


Fig.3 Number of words with vowel in the beginning of words

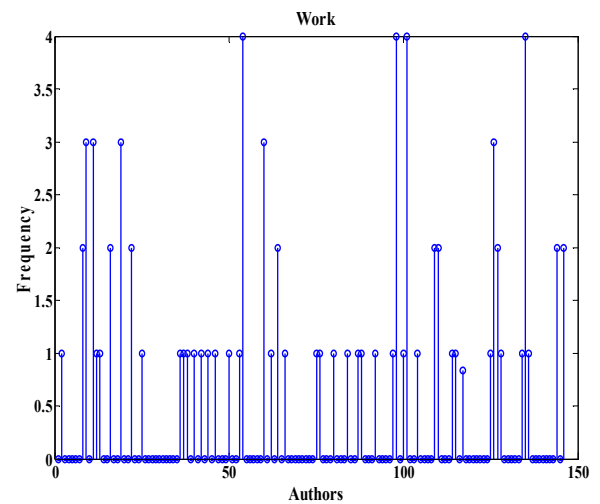


Fig4. Work words for each author

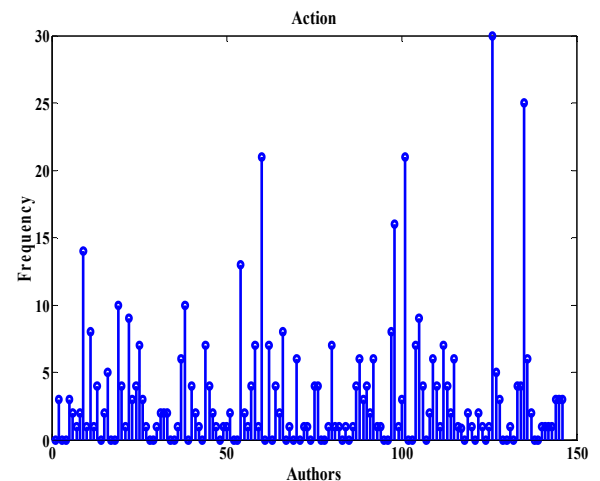


Fig.5 Action words for each author

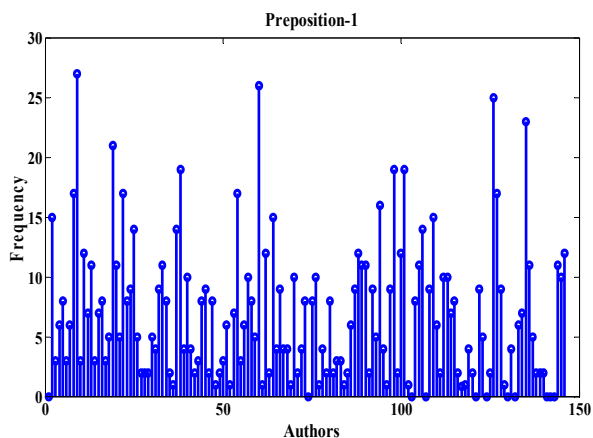


Fig.6 Preposition 1 for each author

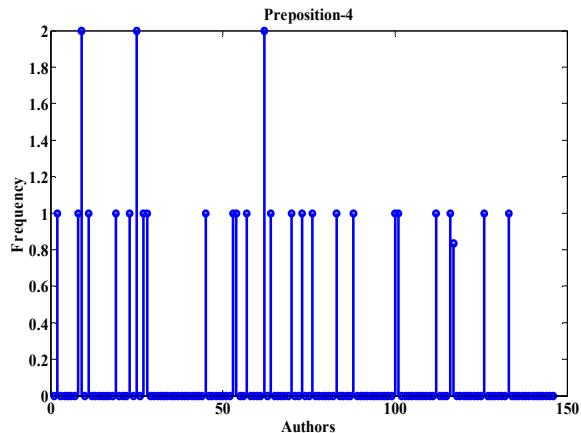


Fig.9 Preposition 4 for each author

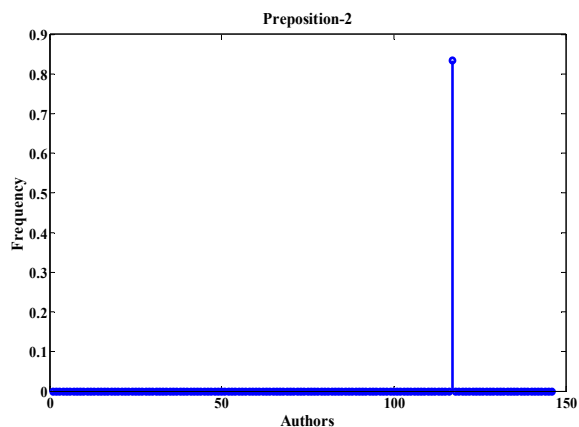


Fig.7 Preposition 2 for each author

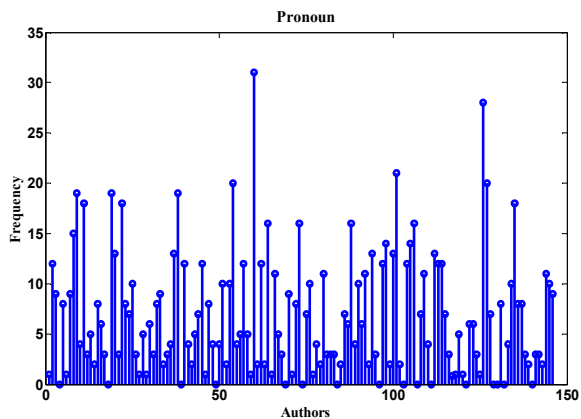


Fig.10 Pronoun for each author

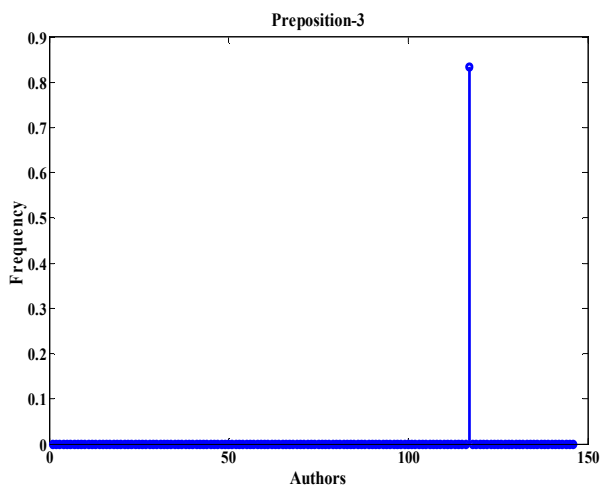


Fig.8 Preposition 3 for each author

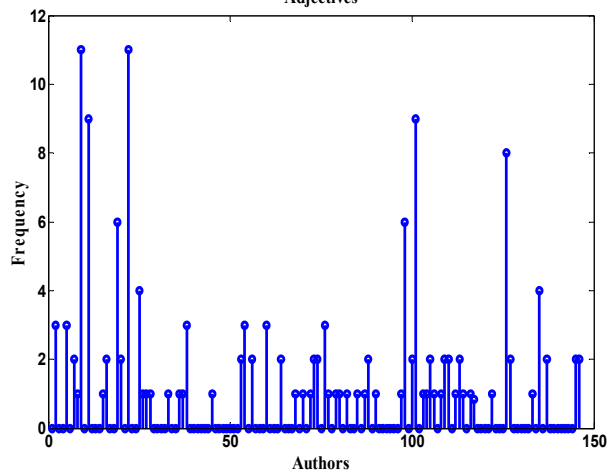


Fig.11 Adjectives for each author

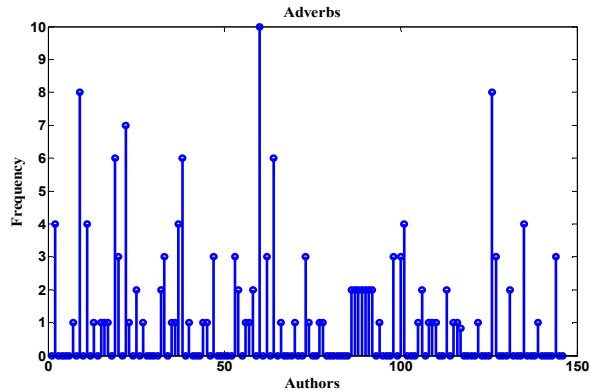


Fig.12 Adverbs for each author

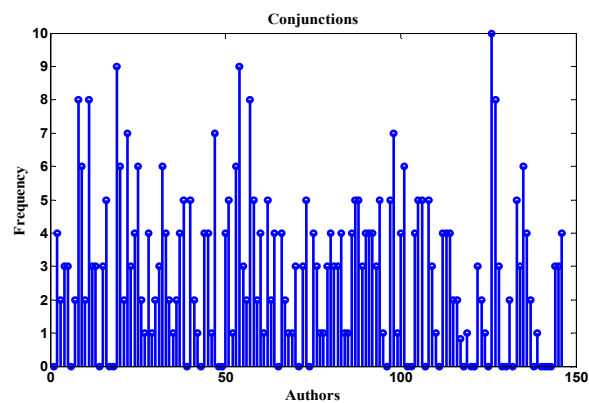


Fig.13 Conjunctions for each author

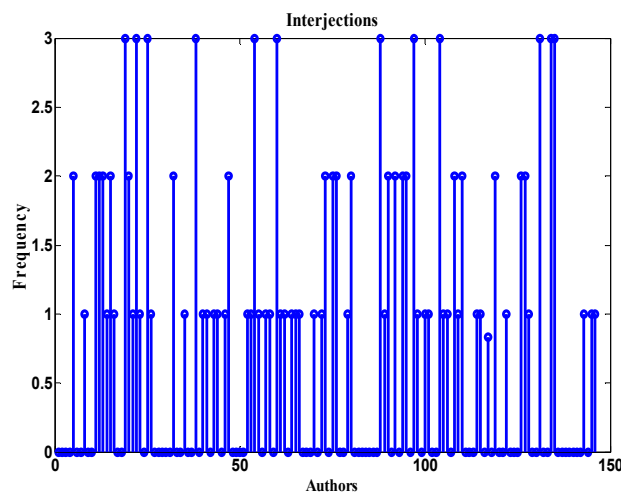


Fig.14 Interjections for each author

We use the following algorithm for email identification by Neural Network training and testing: Find the number of words and their number of occurrences (frequencies) an email and all the emails of an author. Similarly, find the number of words and

their frequencies of all emails of all authors by using the filtering words available in Table 1 -3.

Create a matrix with rows equivalent to total number of unique words considering all emails of all authors. The number of columns is equivalent to number authors.

Based on the dictionary of words obtained from all the emails, fill up a column of the zero matrix based on the availability of the words in a document with their frequencies. Each column will be treated as a pattern for training. A labeling is done for each pattern.

Train the RBF network with patterns considered for training. A final weight matrix is obtained which is further used to test the incoming mails that belongs to existing authors else, the mail can belong to some other person other than these existing authors considered in this experiment.

V. RESULTS AND DISCUSSIONS

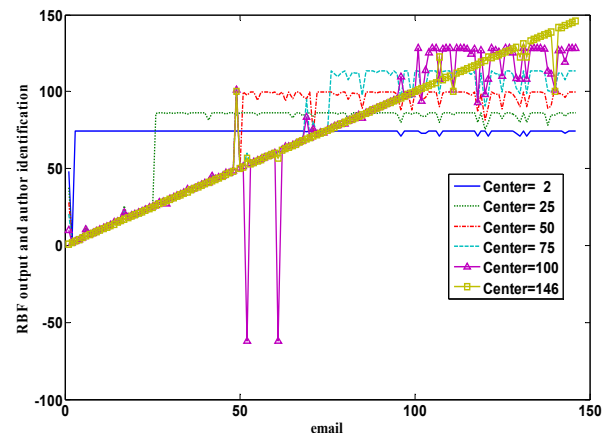


Fig.15 Performance of RBF center selection

The figure 15 presents the performance of RBF in training the patterns. When the number of centers used is less than 50% of the total number of input patterns, the performance of author identification is minimum. As the number of centers increase, the author identification increases. The legend shows the number of centers. Figure 16 presents the performance of the RBF. In this plot, output obtained from RBF overlaps target outputs. The plot emails versus author identification. With 146 centers, the RBF identifies maximum number of authors.

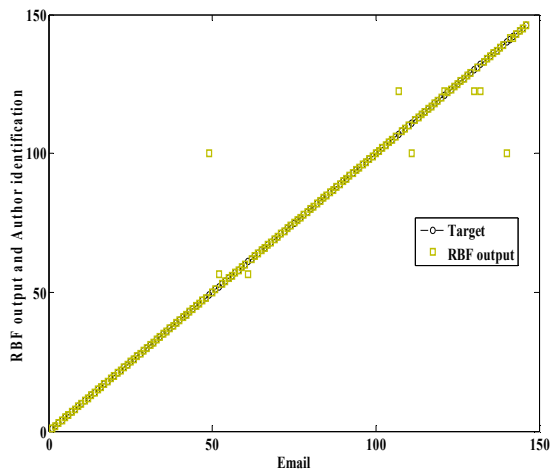


Fig.16 Performance of RBF

This work has presented a novel method of identifying email authorship using RBF patterns of training data have been collected by averaging the frequencies of words of each person and fixing a target value for the person. Testing pattern has been created by modifying the existing contents of an email. A new word has been considered during testing. If the new word does not fit into the patterns used for training, then the word is excluded during testing. As we are unaware to which author the email belongs, now all the training patterns are treated as test patterns after adding the frequencies of the new mail. As only 146 authors are considered, 146 outputs are obtained after testing.

Receiver operating characteristics (ROC) of the authorship identification reveals the following analysis.

Is the author correctly identified of a different document that belongs to this author which is True positive?

Is the author wrongly classified that the document does not belong to him and belongs to some other person or the document does not belong to any one of the ten authors under experiment (False positive)

Is the document that belongs to some other author not considered in this experiment is treated as document of one of the ten authors (False negative).

Is the document considered from outside the training group belongs to same group and not the authors considered in this experiment. (True negative).

Table 6 presents the confusion matrix values and the ROC values. The author emails have been

considered that belong to the training group and that do not belong to training group. All the emails that belong to (sent / sent_mail) folders are used for training. The emails of the remaining folders of all authors have been considered for testing. The performance of RBF has been calculated using confusion matrix. The plot (Figure 17) indicates that the proposed RBF system suits the author identification from given emails. This is inferred from the points obtained above the diagonal of the ROC curve.

TABLE 6 CONFUSION MATRIX FOR RECEIVER
OPERATING CHARACTERISTICS

Instances	True Positive	False Negative	Sensitivity	False Positive	True Negative	Specificity	True Positive Rate
1	80	20	0.80	10	40	0.80	0.20
2	82	18	0.82	8	42	0.84	0.16
3	90	10	0.90	5	45	0.90	0.10
4	85	15	0.85	7	43	0.86	0.14
5	92	8	0.92	8	42	0.84	0.16
Sensitivity=True Positive Rate=True Positive/Total words True Positive Rate=1-Specificity							

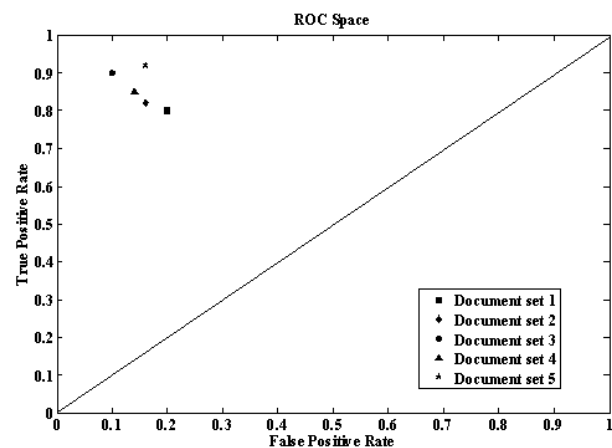


Fig.17 Receiver Operating Characteristics

VI. CONCLUSION

The proposed RBF has been used for author identification of emails. Different RBF centers and their effectiveness in author identification are presented. The receiver operating characteristics curve has been presented and it shows the proposed RBF network performance is acceptable. As a further work, the huge amount of words can be meaningfully filtered that are more specific to an author and that can be further used for author identification.

REFERENCES

1. Abbasi A. And Chen H, "Applying Authorship Analysis to Extremist-Group Web Forum Messages" IEEE INTELLIGENT SYSTEMS, pp. 67-75, 2005.
2. David Madigan, Alexander Genkin, David Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye, "Author Identification on the Large Scale", *Proc. of The Meeting Of The Classification Society of North America*, 2005.
3. Diederich, J., and Chen, H. 2008. Writeprints, "A stylometric approach to identity-level identification and similarity detection", ACM Transactions on Information Systems (26:2), pp. 7.
4. Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003), "Authorship Attribution with Support Vector Machines", *Applied Intelligence* 19(1), pp. 109-123.
5. Goodman R., Hahn M., Marella M., Ojar C., And Westcott S., "The Use Of Stylometry For Email Author Identification: A Feasibility Study", *Proc. Student/Faculty Research Day, CSIS, Pace University, White Plains, NY*, pp.1-7, May 2007.
6. Koppel, M., Schler, J., Argamon, S. and Messeri, E., "Authorship Attribution with Thousands of Candidate Authors", in *Proc. 29th ACM SIGIR Conference on Research & Development on Information Retrieval*, 2006.
8. Moshe Koppel, Shlomo Argamon, And Anat Rachel Shimoni, "Automatically Categorizing Written Texts By Author Gender", *Literary And Linguistic Computation*. 17(4):pp.401-412, 2002.
9. Pavelec, D., Justino, E., And Oliveira, L. S., "Author Identification Using Stylometric Features", *Inteligencia Artificial* (11:36), pp. 59-65, 2007.
10. Peng, F., Schuurmans, D., Wang, S., "Augumenting Naive Bayes Text Classifier With Statistical Language Models", *Information Retrieval*, 7 (3-4), Pp. 317 - 345, 2004.
11. Zheng R., Li J., Chen H., Huang Z., "A Framework For Authorship Identification Of Online Messages: Writing-Style Features And Classification Techniques", *Journal Of The American Society For Information Science And Technology* 57(3):378-93, 2006.

AUTHORS PROFILE



A. PANDIAN received his B.Sc., and MCA degree from Bharathidasan University, Tiruchi. He received his M.Tech degree from Punjabi University, Patiala, Punjab and M.Phil. degree from Periyar University, Salem. He is doing Ph.D. (Computer Science & Engineering) in SRM University, Chennai. He has over fourteen years of experience in teaching. He is working as Assistant Professor (Sr.G) in the Department of MCA, SRM University, Chennai. His areas of interest are text processing, information retrieval and machine learning. He is a member of ISTE and ISC.



Dr. M. Abdul Karim Sadiq holds Ph.D. in Computer Science & Engineering from Indian Institute of Technology, Madras. He has over fourteen years of experience in software development, research, management and teaching. His areas of interest are text processing, information retrieval and machine learning. Having published papers in many international conferences and refereed journals of repute, he filed a patent in the United States Patent and Trademark Office. He is an associate editor of *Soft Computing Applications in Business*, Springer. Moreover, he organized certain international conferences and is on the program committees. He has been awarded a star performer in the software industry.

Performance of Iterative Concatenated Codes with GMSK over Fading Channels

Labib Francis Gergis
Misr Academy for Engineering and Technology
Mansoura, Egypt
IACSIT Senior Member, IAENG Member

Abstract-Concatenated continuous phase modulation (CCPM) facilitates powerful error correction. CPM also has the advantage of being bandwidth efficient and compatible with non-linear amplifiers. Bandwidth efficient concatenated coded modulation schemes were designed for communication over Additive White Gaussian noise (AWGN), and Rayleigh fading channels. Analytical bounds on the performance of Serial (SCCC), and Parallel convolutional concatenated codes (PCCC) were derived as a base of comparison with the third category known as hybrid concatenated convolution codes scheme (HCCC). An upper bound to the soft-input, soft-output (SISO) maximum a posteriori (MAP) decoding algorithm applied to CC's of the three schemes was obtained. Design rules for the parallel, outer, and inner codes that maximize the interleaver's gain were discussed. Finally, a low complexity iterative decoding algorithm that yields a better performance is proposed.

key words: Concatenated codes, continuous phase modulation, GMSK, uniform interleaved coding, convolutional coding, iterative decoding

I. INTRODUCTION

The channel capacity unfortunately only states what data rate is theoretically possible to achieve, but it does not say what codes to use in order to achieve an arbitrary low bit error rate (BER) for this data rate. Therefore, there has traditionally been a gap between the theoretical limit and the

achievable data rates obtained using codes of a manageable decoding complexity.

However, a novel approach to error control coding revolutionized the area of coding theory. The so-called turbo codes [1,2], almost completely closed the gap between the theoretical limit and the data rate obtained using practical implementations. Turbo codes are based on concatenated codes (CC's) separated by interleavers. The concatenated code can be decoded using a low-complexity iterative decoding algorithm [3]. Given certain conditions, the iterative decoding algorithm performs close to the fundamental Shannon capacity. In general, concatenated coding provides longer codes yielding significant performance improvements at reasonable complexity investments. The overall decoding complexity of the iterative decoding algorithm for a concatenated code is lower than that required for a single code of the corresponding performance.

The parallel, serial, and hybrid concatenation of codes are well established as a practical means of achieving excellent performance. Interest in code concatenation has been renewed with the introduction of turbo codes [4,5,6,7]. These codes perform well and yet have a low overall decoding complexity.

CPM is a form of constant-envelope digital modulation and therefore of interest for use with nonlinear and/or fading channels. The inherent bandwidth- and energy efficiency makes CPM a very attractive modulation scheme [8]. Furthermore, CPM signals have good spectral properties due to their phase continuity. Besides providing spectral economy, CPM schemes exhibit a "coding gain" when compared to PSK modulation.

This "coding gain" is due to the memory that is introduced by the phase-shaping filter and the decoder can exploit this. CPM modulation exhibits memory that resembles in many ways how a convolutionally encoded data sequence exhibits memory - in both cases, a "trellis" can be used to display the possible output signals (this is why convolutional encoders are used with CPM in this paper).

This paper is organized as follows. Section II briefly describes continuous phase modulation, using Gaussian minimum shift keying GMSK, and how it can be separated into a finite-state machine and a memoryless signal mapper. Section III describes in details the system model and encoder structure of serial, parallel, and hybrid concatenated codes.

Section IV derives analytical upper bounds to the bit-error probability of the three concatenated codes using the concept of uniform interleavers that decouples the output of the outer encoder from the input of the inner encoder, from the knowledge of the input-output weight coefficients (IOWC), $A_{w,h}^c$ for CC's. $A_{w,h}^c$ is represented related to the type of concatenation. The choice of decoding algorithm and number of decoder iterations is described in section V. Factors that affect the performance are discussed in section VI. Finally conclusion results for some examples described in section IV have been considered in section VII.

II. GMSK SYSTEM MODEL

Gaussian minimum-shift keying is a special case of a more general class of modulation schemes known as continuous phase modulation (CPM). In CPM schemes, the signal envelope is kept constant and the phase varies in a continuous manner. This ensures that CPM signals do not have the high-frequency components associated with sharp changes in the signal envelope and allows for more compact spectra. CPM signal $s(t)$ can be written [8,9]

$$S(t) = (\sqrt{2E_b/T}) \cos [2\pi f_c t + \Phi(t+\alpha)] \quad (1)$$

where E_b is the energy per symbol interval, T is the duration of the symbol interval, f_c is the carrier frequency, and $\Phi(t+\alpha)$ is the "phase function" responsible for mapping the input sequence to a corresponding phase waveform.

The term $\alpha = \{a_i\}$ is the input sequence taken from the M-ary alphabet $\pm 1, \pm 3, \dots, \pm M - 1$. For convenience the focus here will be on the binary case, $a_i \in \{\pm 1\}$.

The "continuous phase" constraint in CPM requires that the phase function maintain a continuous amplitude. In general the phase function is given by

$$\Phi(t+\alpha) = 2\pi \sum_{n=0}^N a_n \delta_n g(t - nT) \quad (2)$$

where δ is the modulation index, and $g(t)$ is the phase pulse. The phase pulse $g(t)$ is typically specified in terms of a normalized, time-limited frequency pulse $f(t)$ of duration LT such that:

$$g(t) = \begin{cases} 0 & ; \quad \text{if } t < 0 \\ \int_0^t f(\tau) d\tau & ; \quad \text{if } 0 < t < LT \\ 1/2 & ; \quad \text{if } t > LT \end{cases} \quad (3)$$

The duration term (LT) is specified in terms of the bit duration T , and identifies the number of bit durations over which the frequency pulse $f(t)$ is non-zero, $\delta = 1/2$, and the frequency pulse is

$$f(t) = (1/2T) \left\{ Q \left[\frac{2\pi B (t - \tau/2)}{\sqrt{\ln 2}} \right] - Q \left[\frac{2\pi B (t + \tau/2)}{\sqrt{\ln 2}} \right] \right\} \quad (4)$$

B is a parameter in GMSK which controls the amount of bandwidth used as well as the severity of the intersymbol interference, the B parameter is expressed in terms of the inverse of the bit duration T .

III. PERFORMANCE ANALYSIS OF COCATENATED CODES

Consider a linear (n,k) code C with code rate $R_c = k/n$ and minimum distance d_{min} . An upper bound on the bit-error rate [BER] of the code C over memoryless binary-input channels, with coherent detection, using maximum likelihood decoding, can be obtained as [4]

$$BER \leq \sum_{h=d_{min}}^n \sum_{w=1}^k (w/k) A_{w,h}^c D(R_c E_b / N_o, h) \quad (5)$$

where E_b/N_o is the signal-to-noise ratio per bit, and $A_{w,h}^c$ for the code C represents the number of codewords of the code with output weight h associated with an input sequence of weight w . $A_{w,h}^c$ is the input-output weight coefficient (IOWC). The function $D(.)$ represents the pairwise error probability which is a monotonic decreasing function of the signal to noise ratio and the output weight h . For AWGN channels we have $D(R_c E_b / N_o, h) = Q(\sqrt{2R_c h E_b / N_o})$.

For fading channels, assuming coherent detection, and perfect Channel State information (CSI), the fading samples μ_i are i.i.d. random variables with Rayleigh density of the form $f(\mu) = 2\mu e^{-\mu^2}$. The conditional pairwise error probability is given by

$$D(R_c E_b / N_o, h | \mu) = Q\left(\sqrt{\frac{h}{2R_c h E_b / N_o} \sum_{i=1}^h \mu_i^2}\right) \quad (6)$$

where Q function can be defined as

$$Q(x) \leq (1/2) e^{-x^2/2} \quad (7)$$

By averaging the conditional bit error rate over fading using (5), (6), and (7). The upper bound for BER is represented by

$$BER \leq 0.5 \sum_{h=h_{min}}^n \sum_{w=1}^k (w/k) A_{w,h}^c \cdot [1/(1+R_c E_b / N_o)]^h \quad (8)$$

It is clear from equation (8) that BER depends on major factors like signal-to-noise ratio per bit, and the input-output weight coefficients (IOWC), $A_{w,h}^c$ for the code, $A_{w,h}^c$ is represented related to the type of concatenation.

The average IOWC for λ concatenated codes with $\lambda-1$ interleavers can be obtained by averaging (5) over all over all possible interleavers. This average is obtained by replacing the actual i^{th} interleaver ($i = 1, 2, \dots, \lambda-1$), that performs a permutation of the N_i input bits, with an abstract interleaver called uniform interleaver defined as a probabilistic device that maps a given input word of weight w into all distinct $\binom{N_i}{w}$

permutations of it with equal probability $\psi = 1 / \binom{N_i}{w}$

IV. DESIGN OF CONCATENATED CODES

Concatenated codes represent a more recent development in the coding research field [1], which has risen a large interest in the coding community.

IV. 1. Design of Parallel Concatenated Convolutional Codes (PCCC)

The first type of concatenated codes is *parallel concatenated convolutional codes* (PCCC) whose encoder is formed by two (or more) *constituent* systematic encoders joined through one or more interleavers. The input information bits feed the first encoder and, after having been scrambled by the interleaver, they enter the second encoder. A codeword of a parallel concatenated code consists of the input bits to the first encoder followed by the parity check bits of both encoders. As shown in Fig. 1, the structure of PCCC consists of convolutional code C_1 with rate $R_p^1 = p/q_1$, and convolutional code C_2 with rate $R_p^2 = p/q_2$, where the constituent code inputs are joined by an interleaver of length N , generating a PCCC, C_p , with total rate R_p . The output codeword length $n = n_1 + n_2$ [4].

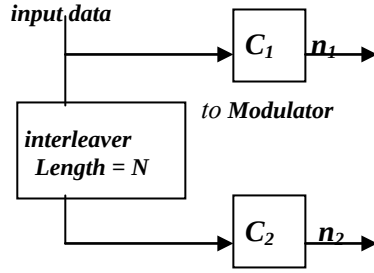


Fig. 1. Parallel Concatenated Convolutional Code (PCCC)

The input–output weight coefficients (IOWC), $A_{w,h}^{cp}$ for PCCC can be defined as [3]

$$A_{w,h}^{cp} = \sum_{\substack{h_1, h_2 : \\ h_1 + h_2 = h}} A_{w,h_1,h_2}^{cp} = \sum_{\substack{h_1, h_2 : \\ h_1 + h_2 = h}} \frac{A_{w,h_1}^{c_1} \times A_{w,h_2}^{c_2}}{\binom{N}{w}} \quad (9)$$

where A_{w,h_1,h_2}^{cp} is the number of codeword of the PCCC with output weights h_1 , and h_2 associated with an input sequence of weight w .

Let $A_{w,h,j}^c$ be the IOWC given that the convolutional code generates j error events with total input w , and output weight h . The $A_{w,h,j}^c$ actually represents the number of sequences of weight h , input weight w , and the number of concatenated error events j without any gap between them, starting at the beginning of the code. For N much larger than the memory of the convolutional code, the coefficient of the equivalent code can be approximated by

$$A_{w,h}^c \approx \sum_{j=1}^{n_M} \binom{N/p}{j} A_{w,h,j}^c \quad (10)$$

where n_M , the largest number of error events concatenated in a codeword of weight h and generated by a weight w input sequence, is a function of h and w that depends on the decoder.

$$\binom{N}{j} \approx N^j / j! \quad (11)$$

Substitution of this approximation in (10) Yields

$$A_{w,h}^c \approx \sum_{j=1}^{n_M} (N^j / j! p!) A_{w,h,j}^c \quad (12)$$

Inserting (12) into (9), we obtain the input–output weight coefficients (IOWC), $A_{w,h}^{cp}$ for PCCC as [4]

$$\begin{aligned} A_{w,h}^{cp} &\approx \sum_{n_1=1}^{n_{max}} \sum_{n_2=1}^{n_{max}} \frac{\binom{N/p}{n_1} \binom{N/p}{n_2}}{\binom{N}{w}} \\ &\quad \cdot A_{w,h,n_1}^{c_1} A_{w,h,n_2}^{c_2} \\ &\approx \sum_{n_1=1}^{n_{max}} \sum_{n_2=1}^{n_{max}} \frac{w!}{p^{n_1+n_2} n_1! \cdot n_2!} \\ &\quad \cdot N^{n_1+n_2-w} \cdot A_{w,h,n_1}^{c_1} A_{w,h,n_2}^{c_2} \end{aligned} \quad (13)$$

IV. 2. Design of Serially Concatenated Convolutional Codes (SCCC)

Another equally powerful code configuration with comparable performance to parallel concatenated codes is serially concatenated convolutional codes (SCCC).

The structure of a serially concatenated convolutional code (SCCC) is shown in Fig. 2. It refers to the case of two convolutional CCs, the outer code C_o with rate $R_c^o = k/p$, and the inner code C_i with rate $R_c^i = p/n$, joined by an interleaver length N bits, generating an SCCC with rate $R_c = k/n$.

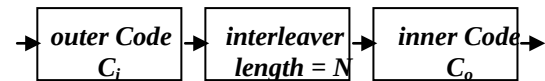


Fig. 2. Serially Concatenated Convolutional Code (SCCC)

From the knowledge of the IOWC of outer and inner codes, which called $A^{co}(w,L)$ and $A^{ci}(l,H)$. Exploit the properties of the uniform interleaver, which transforms a codeword of weight l at the output of the outer encoder into all its distinct $\binom{N}{l}$ permutations.

As a consequence, each codeword of the outer code C_o of weight l , through the action of the uniform interleaver, enters the inner encoder generating $\binom{N}{l}$ codewords of the inner code C_i .

Thus, the IOWC of the SCCC scheme, $A_{w,h}^{cs}$, of codewords with weight h associated with an input word of weight w is given by

$$A_{w,h}^{cs} = \sum_{l=0}^N \frac{A_{w,l}^{co} \times A_{l,h}^{ci}}{\binom{N}{l}} \quad (14)$$

Using the previous result of (12) with $j=n^i$ for the inner code, and the analogous one, $j=n^o$, for the outer code.

$$A_{w,l}^{co} \approx \sum_{n^o=1}^{n_o^m} \binom{N/p}{n^o} A_{w,l,n^o}^o \quad (15)$$

Substituting (15) into (14) defines $A_{w,h}^{cs}$ for SCCC in the form

$$A_{w,h}^{cs} \approx \sum_{l=d_f^o}^N \sum_{n^o=1}^{n_{max}} \sum_{n^i=1}^{n_{max}} \frac{\binom{N/p}{n_1} \binom{N/p}{n_2}}{\binom{N}{l}} \cdot A_{w,l,n^o}^o \cdot A_{l,h,n^i}^i \quad (16)$$

where d_f^o is the free distance of the outer code. Inserting the approximation (11) in (16) yields

$$A_{w,h}^{cs} \approx \sum_{l=d_f^o}^N \sum_{n^o=1}^{n_o^m} \sum_{n^i=1}^{n_i^m} \frac{l!}{p^{n^o+n^i} \cdot n^o! \cdot n^i!} \cdot N^{n^o+n^i} \cdot A_{w,l,n^o}^o \cdot A_{l,h,n^i}^i \quad (17)$$

IV. 3. Design of Hybrid Concatenated Convolutional Codes (HCCC)

The structure of a hybrid concatenated convolutional code is shown in Fig. 3. It is

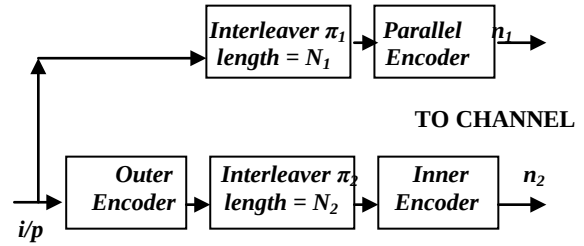


Fig. 3. Hybrid Concatenated Convolutional Code (HCCC)

composed of three concatenated codes, the parallel code C_p with rate $R_c^p = k_p/n_p$, the outer code C_o with rate $R_c^o = k_o/p_o$, and the inner code C_i with rate $R_c^i = p_i/n_i$, with two interleavers N_1 and N_2 bits long. Generating an HCCC C_H with overall rate R_H . For simplicity, assuming $k_p = k_o$ and $p_o = p_i = p$, then $R_H = k_o / (n_p + n_i)$.

Since the HCCC has two outputs, the upper bounds on the bit error probability in (8) can be modified to

$$BER \leq \sum_{h=h^p}^{n_1} \sum_{h=h^i}^{n_2} \sum_{w=w_m}^k (w/k) \cdot A_{w,h_1,h_2}^{CH} \cdot Q \left[\sqrt{2R_H (h_1 + h_2) (E_b / N_o)} \right] \quad (18)$$

where A_{w,h_1,h_2}^{CH} for the HCCC code represents the number of codewords with output weight h_1 for the parallel code and output weight h_2 for the inner code, associated with an input sequence of weight w , A_{w,h_1,h_2}^{CH} is the IOWC for HCCC, w_m is the minimum weight of an input sequence generating the error events of the parallel code and the outer code, h_p is the minimum weight of the codewords of C_p , and h_i is the minimum weight of the codewords of C_i .

With knowledge of the IOWC A_{w,h_1}^{Cp} for the constituent parallel code, the IOWC $A_{w,l}^{Co}$ for the constituent outer code, and the IOWC A_{l,h_2}^{Ci} for the constituent inner code,

using the concept of the uniform interleaver, the A_{w,h_1,h_2}^{CH} for HCCC can be obtained.

$$A_{w,h_1,h_2}^{CH} = \sum_{l=0}^{N_2} \frac{A_{w,h_1}^{cp} \times A_{w,l}^{co} \times A_{l,h_2}^{ci}}{\binom{N_1}{w} \binom{N_1}{l}} \quad (19)$$

$A_{w,h}^{CH}$ can be obtained by summing A_{w,h_1,h_2}^{CH} overall h_1 , and h_2 such that $h_1+h_2=h$. $A_{w,l}^{Co}$ is the number of codewords of C_o of weight l given by the input sequences of weight w . Analogous definitions apply for A_{w,h_1}^{cp} and A_{l,h_2}^{ci} .

V. EXAMPLES CONFIRMING THE DESIGN OF CONCATENATED CODES RULES

To obtain the design rules obtained asymptotically, for different signal-to-noise ratios and large interleaver lengths, N , the upper bounds for (8) to BER for several types of the concatenated codes were evaluated, with different interleaver lengths, and compare their performances with those predicted by the design guidelines.

1. Parallel Concatenated Convolutional Codes (PCCC)

Consider a PCCC with overall rate = $1/3$, formed by two convolutional codes, C_1 , and C_2 , have equal rate = $1/2$, linked through an uniform interleaver with length N , and whose encoder is shown in Fig. 1. We have constructed different PCCCs through interleavers of various lengths, and passed through the previous steps to evaluate their performance with GMSK modulator.

Upper bounds to the error probability based on the union bound described in(8), and (13) present a divergence at low values of signal-to-noise ratio. Fig. 4. shows the bit error probability of a PCCC over AWGN,

and Rayleigh fading channels with GMSK modulation scheme, using interleaver of lengths $N = 100, 1000$, and 2000 bits.

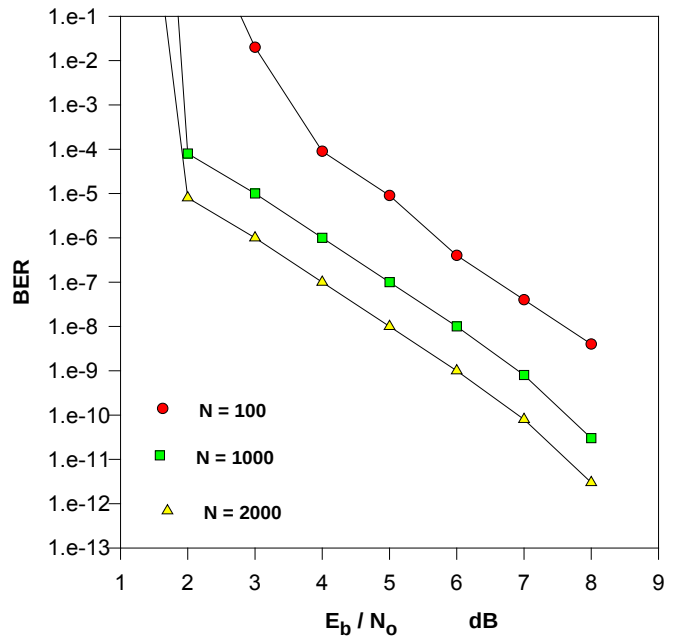


Fig. 4. Analytical bounds for PCCC with GMSK Modulation Scheme through different interleaver lengths

2. Serially Concatenated Convolutional Codes (SCCC)

Consider a rate $1/3$ SCCC using as outer code a convolutional encoder C_o with rate $R_c^o = 1/2$, and the inner code C_i with rate $R_c^i = 2/3$, joined by an uniform interleaver of length $N = 100, 1000$, and 2000 bits, as shown in Fig. 2. Using the previously analysis for SCCC that defined in (8), and (17), we obtained the bit-error probability bounds illustrated in Fig. 5. The performance was obtained over AWGN, and Rayleigh fading channels with GMSK modulation scheme.

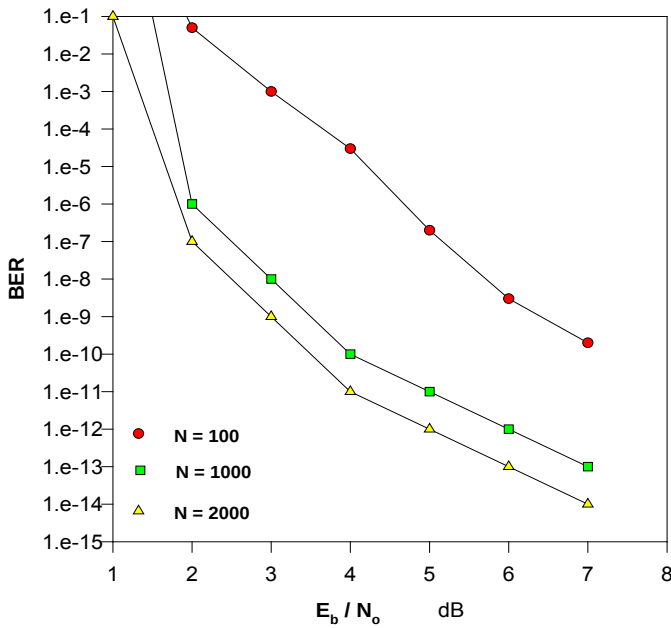


Fig. 5. Analytical bounds for SCCC with GMSK Modulation Scheme through different interleaver lengths

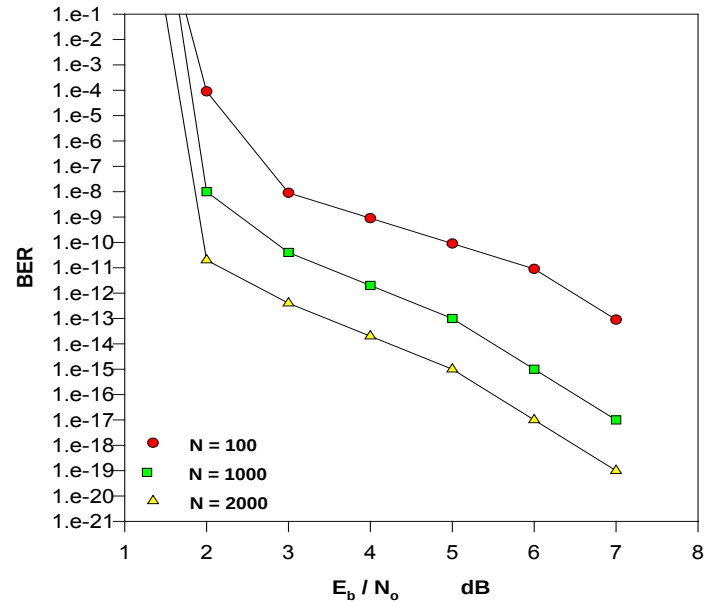


Fig. 6. Analytical bounds for HCCC with GMSK Modulation Scheme through different interleaver lengths

3. Hybrid Concatenated Convolutional Codes (HCCC)

Consider a rate $1/4$ HCCC formed by a parallel systematic convolutional code with rate $1/2$, an outer four convolutional code with rate $1/2$, and an inner convolutional code with rate $2/3$, joined by two uniform interleavers of length $N_1 = N$ and $N_2 = 2N$, where $N=100, 1000$, and 2000 bits, as shown in Fig. 3. Using (8), and (18) we have obtained the bit error probability curves over AWGN, and Rayleigh fading channels with GMSK modulation scheme, shown in Fig. 6.

VI. ITERATIVE DECODING OF CONCATENATED CODES

Maximum-likelihood (ML) decoding of SCCC, PCCC, and HCCC with large N is an almost complex and impossible achievement. To acquire a practical significance, an iterative algorithm consists of a soft-input, soft-output (SISO) maximum a posteriori (MAP) decoding algorithm applied to CC's [10], [11], and [12]. A functional diagram of the iterative decoding algorithm for PCCC, SCCC, and HCCC are illustrated in figures 7, 8, and 9, respectively.

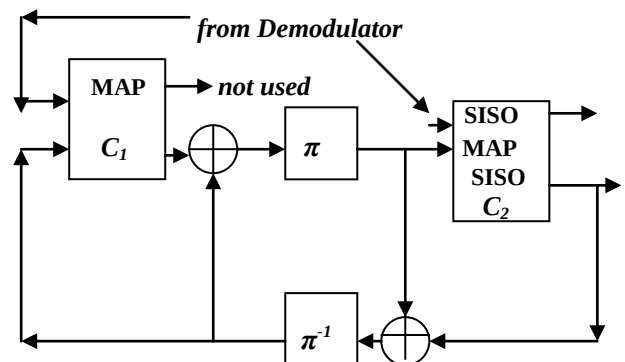


Fig. 7. Iterative decoding algorithm for PCCC

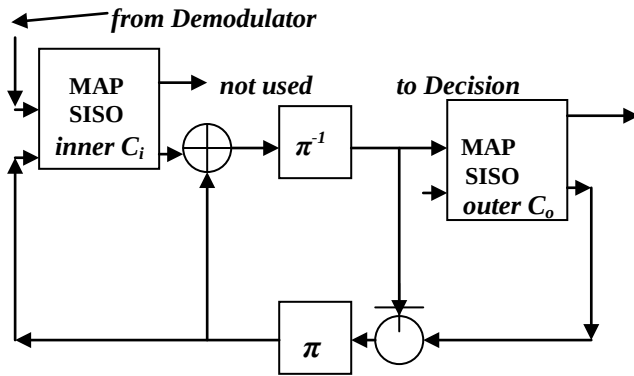


Fig. 8. Iterative decoding algorithm for SCCC

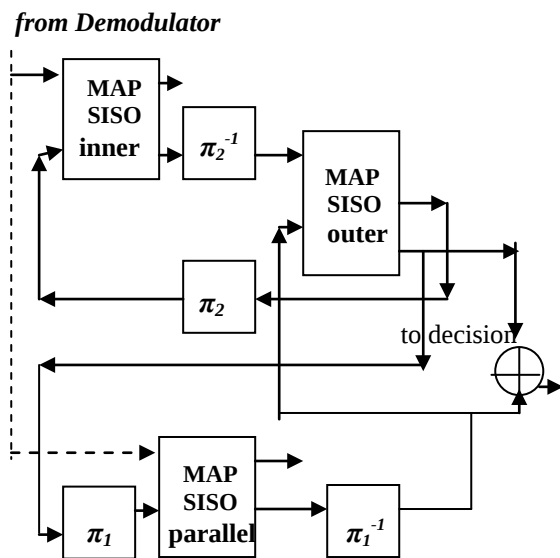


Fig. 9. Iterative decoding algorithm for HCCC

VI. THE EFFECT OF VARIOUS PARAMETERS ON THE PERFORMANCE OF CONCATENATED CODES

The performances of concatenated codes were evaluated and analyzed in the previous sections. There are many parameters which affect the performance of CC's when decoded with iterative decoder over AWGN and Rayleigh fading channels. It is shown, briefly, the effect of the interleavers lengths, and the number of decoding iterations.

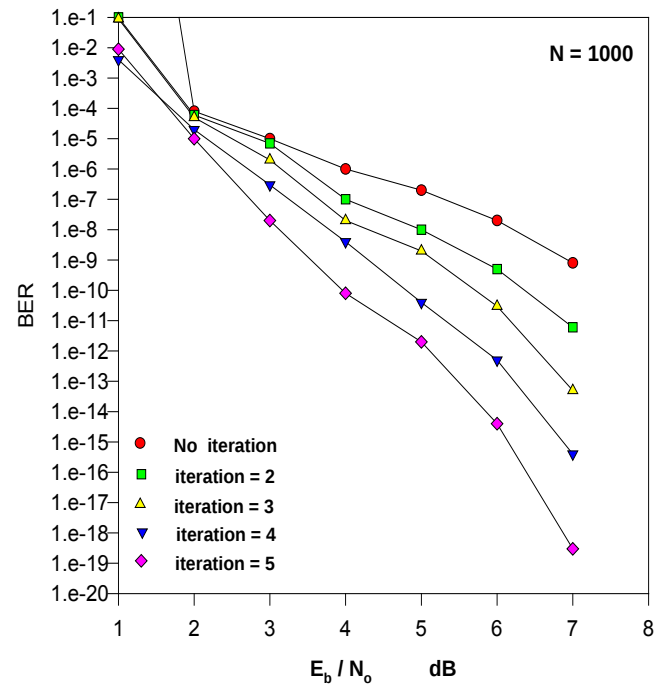


Fig. 10. Analysis of Iterative decoding algorithm for 1/3 PCCC with different No of iterations

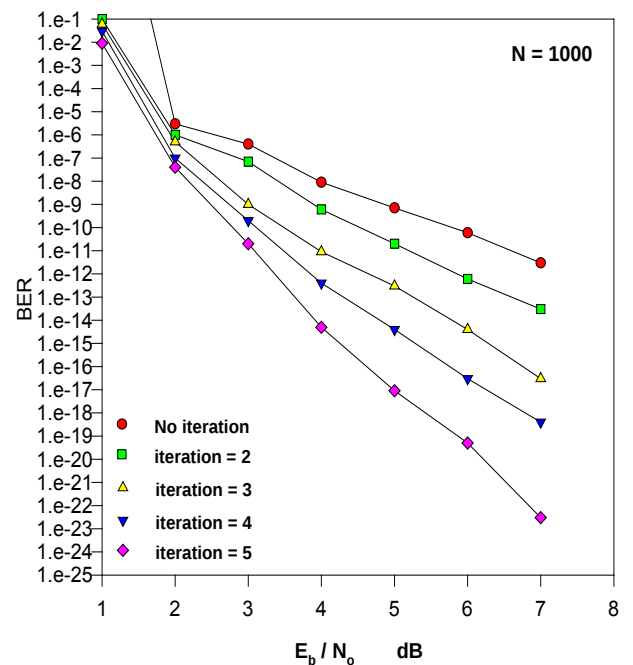


Fig. 11. Analysis of Iterative decoding algorithm for 1/3 SCCC with different No of iterations

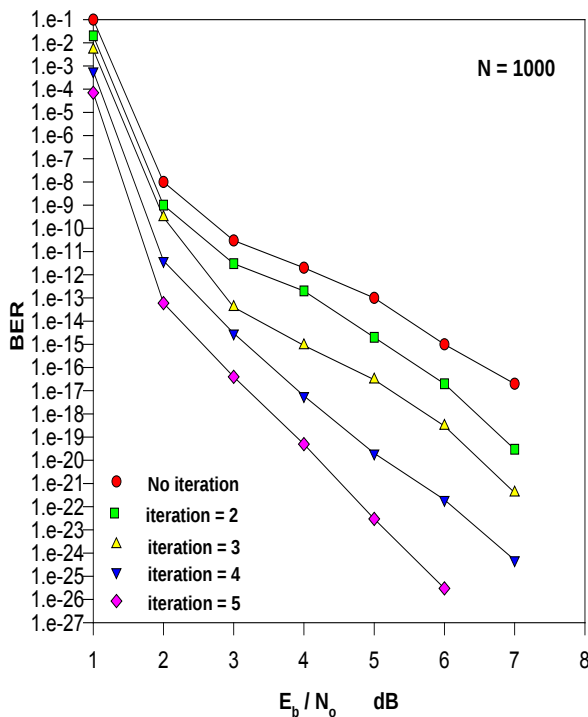


Fig. 12. Analysis of Iterative decoding algorithm for 1/4 HCCC with different No of iterations

A. The effect of interleaver length

It is well known that a good interleaving affects the CC's error performance considerably. Figures 4, 5, and 6 represent the BER of PCCC, SCCC, and HCCC, respectively, versus the interleaving length, N . From these figures, it is shown that BER improve with increasing the length of interleaver.

B. The effect of number of decoder iterations

The choice of decoding algorithm and number of decoder iterations also influences performance. A functional diagrams of the iterative decoding algorithm for CC's were presented in figures 7, 8, and 9 for PCCC, SCCC, and HCCC, respectively. It could be observed, from figures 10, 11, and 12, that the slope of curves and coding gain are improved by increasing the number of iterations.

VII . CONCLUSIONS

A construction of concatenated codes CC's have presented and constructed in this paper with three main basic schemes: PCCC, SCCC, and HCCC, over AWGN and Rayleigh fading channels. The effects of various parameters on the performance of CC's, using an upper bound to the soft-input, soft-output (SISO) maximum a posteriori (MAP) decoding algorithm are investigated. These parameters are : the interleaver length, and the number of iterations. The analytical results showed that coding gain was improved by increasing the interleaver length, and the number of iterations.

REFERENCES

- [1] N. Sven, " Coding and Modulation for Spectral Efficient Transmission ", PhD dissertation .. Institute of Telecommunications of the University of Stuttgart, 2010.
- [2] Z. Chance, and D. Love, " Concatenated Coding for the AWGN Channel with Noise Feedback ", School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, 2010.
- [3] E. Lo, and K. Lataief, " Cooperative Concatenated Coding for Wireless Systems", Proceeding of IEEE Communication Society, 2007.
- [4] B. Cristea, " Viterbi Algorithm for Iterative Decoding of Parallel Concatenated Convolutional Codes ", 18th European Signal Processing (EUSIPCO), Aalborg, Denmark, August 2010.
- [5] R. Maunder, and L. Hanzo, " Iterative Decoding Convergence and Termination of Serially Concatenated Codes ", IEEE Transactions on Vehicular Technology, VOL. 59, NO. 1, January 2010.

- [6] C. Koller, A. Amat, Kliewer, F. Vatta, and D. Costello, " Hybrid Concatenated Codes with Asymptotically Good Distance Growth ", Proceeding of 5th International Symposium on Turbo Codes and Related Topics, 2008.
- [7] N. Wahab, I. Ibrahim, S. Sarnin, and N. Isa, Performance of Hybrid Automatic Repeat Request Scheme with Turbo Codes ", Proceeding of the International Multi Conference of Engineering and Computer Scientists (IMECS), VOL. II, Hong Kong, March 2010.
- [8] K. Damodaran, " Serially Concatenated Coded Continuous Phase Modulation for Aeronautical Telemetry ", Master of Science thesis, Department of Electrical Engineering, Faculty of the Graduate School of the University of Kansas. August 2008.
- [9] A. Perotti, P. Remlein, and S. Benedetto , " Adaptive Coded Continuous-Phase Modulations for Frequency-Division Multiuser Systems ", Advances in Electronics and Telecommunications, VOL. 1, NO. 1, April 2010.
- [10] M. Lahmer, and M. Belkasmi, " A New Iterative Threshold Decoding Algorithm for one Step Majority Logic Decodable Block Codes ", International Journal of Computer Applications, Volume 7, No. 7, October 2010.
- [11] F. Ayoub, M. Lahmer, M. Belkasmi, and E. Bouyakhf, " Impact of the Decoder Connection Schemes on Iterative Decoding of GPCB Codes ", International Journal of Information and Communication Engineering, VOL. 6, No. 3, 2010.
- [12] A. Farchane, and M. Belkasmi , " Generalized Serially Concatenated Codes; Construction and Iterative Decoding ", International Journal of Mathematical and Computer Sciences, VOL. 6, No. 2, 2010.

Priority Based Mobile Transaction Scheme Using Mobile Agents

J.L.Walter Jeyakumar^{#1}, R.S.Rajesh^{#2}

[#] *Department of Computer Science and Engineering,
Manonmaniam Sundaranar University,*

Tirunelveli, Tamilnadu, INDIA.

¹ walterjeya@hotmail.com

² rs_rajesh@yahoo.co.in

Abstract— We define a priority based mobile transaction scheme in which mobile users can share data stored in the cache of a mobile agent which is a special mobile node for coordinating the sharing process. This framework allows mobile affiliation work group to be formed dynamically with a mobile agent and mobile hosts. Using short range wireless communication technology, mobile users can simultaneously access the data from the cache of the mobile agent. Data Access Manager module at the mobile agent enforces concurrency control using cache invalidation technique. Four levels of priority are assigned to the requesting mobile nodes based on available energy and connectivity. This model supports disconnected mobile computing by allowing mobile agent to move along with the Mobile Hosts. The proposed Transaction frame work has been simulated in J2ME and NS2 and performance of this scheme is compared with existing frame works.

Key words – Transaction, concurrency control, mobile database, cache invalidation, mobility

I. INTRODUCTION

Mobile computing environment consists of Fixed Hosts (FHs), Mobile Hosts (MHs) and Base stations or Mobile Support Stations (MSSs). MH is connected to the Fixed network through MSS via wireless channels. The Geographical area covered by a MSS is called a cell. Mobile Hosts are portable computers which move around in a cell. When a MH enters into a new cell hand-off or hand-over takes place. MH communicates only with the MSS responsible for its cell. Transactions and data management functions are done using the data base servers installed at MSS. In mobile computing, it is necessary that a computation is not disrupted while an MH is not connected. The part of the computation executing on an MH might continue executing concurrently with the rest of the computations while the MH is moving and not connected to the network [8]. With the evolution of PCS(personal Communication System) and GSM(Global System for Mobile communication), advanced wireless communication services are being offered to the mobile users. Mobile Database System is a distributed client/server system based on PCS or GSM in which clients can move around freely while performing their data processing activities in connected or disconnected mode. Frequent disconnections, mobility, limited battery power and resources pose new challenges to mobile computing environment. Frequent aborts

due to disconnection should be minimized in mobile transactions. The low and variable bandwidth of wireless network together with the expensive transmission cost makes bandwidth consumption an important concern [2]. Correctness of transactions executed on both fixed and mobile hosts must be ensured by the operations on shared data. Blocking of mobile transactions due to long disconnection periods should be minimized to reduce communication cost and to increase concurrency [3]. After disconnection, mobile host should be able to process transactions and commit locally. In Mobile computing, there is always a competition for shared data since it provides users with the ability to access information through wireless connections that can be retained even while the user is moving. Further, mobile users are required to share their data with others. This provides the possibility of concurrent access of data by mobile hosts which may result in data inconsistency. Concurrency control methods have been used to control concurrency. Due to limitations and restrictions of wireless communication channels, it is difficult to ensure consistency of data while sharing takes place.

In this paper, we present a priority based mobile transaction frame work that allows mobile users to share data cached in the Mobile Agent which is a special node for coordinating the sharing process. Whenever an MH enters into a Mobile Agent area it can connect and access the data in the cache. But upon update request by a MH, updation is done at the local cache and invalidation report is sent to all the mobile hosts which have already accessed the same data. This will force the mobile hosts to refresh their data values. This framework also provides the provision for transaction update during disconnection. Data Access Manager (DAM) at the Mobile Agent will take care of concurrency control while sharing takes place. Concurrency control is enforced using cache invalidation technique. In order to give priority to the mobile nodes running on low power and with low connectivity, four levels of priority are used. We also take into account the mobility of the Mobile Hosts that has a strong impact on mobile applications [4].

The remaining part of this paper is organized as follows. Section II summarizes the related research. Section III focuses on the Mobile Agent based architecture. Section IV specifies the proposed framework for disconnected mobile computing. Section V gives the performance analysis. In Section VI, the

discussion on the proposed model is presented. Finally, Section VII concludes the paper.

II. RELATED WORK

When simultaneous access to data is made at the server, concurrency control techniques are employed to avoid data inconsistency. Conventional locking based concurrency control methods like centralized Two Phase locking and distributed Two Phase locking are not suitable for mobile environment. In centralized two phase locking scheme [17, 19], where one node is responsible for managing all locking activities, the problem of single point failure cannot be avoided. The distributed two phase locking scheme used in [18], allows all nodes to serve as lock managers. But in the event of data partition, this algorithm could degenerate into a centralized two phase scheme. In conventional locking scheme, the communication overhead that arises due to locking and unlocking requests can create a serious performance problem because of low capacity and limited resources in mobile environment [5]. Moreover, it makes mobile hosts to communicate with the server continuously to obtain and manage locks [6].

The timestamp approach for serializing the execution of concurrent transactions was developed for the purpose of more flexibility, to eliminate the cost of locking, and to cater for distributed database systems [7, 13]. In timestamping, the execution order of concurrent transactions is defined before they begin their execution. The execution order is established by associating a unique timestamp to every transaction. When two transactions conflict over a data item, their timestamps are used to enforce serialization by rolling back one of the conflicting transactions. To exploit the dynamic aspects of two phase locking and the static ordering of timestamping, a number of concurrency control techniques were developed using a combined approach. In mixed approach techniques called Wound-wait and Wait-die [7, 16], the enforcement of mutual exclusion among transactions is carried out using locking while conflicts are resolved using timestamps.

In [14], optimistic concurrency control scheme is used to minimize locking overhead by delaying lock operation until conflicting transactions are ready to commit. They rely on efficiency in the hope that conflicts between transactions will not occur. Without using lock during the execution of the transactions, this scheme promotes deadlock free execution. In optimistic concurrency control with dynamic time stamp adjustment protocol, client side write operations are required. But it may never be executed due to delay in execution of a transaction [1]. In multi version transaction model [9], data is made available as soon as a transaction commits at a mobile host and another transaction can share this data. But data may be locked for a longer time at a mobile host before the lock is released at the database server.

In [11], AVI (Absolute Validity Interval) was introduced for enforcing concurrency control without locking. AVI is the valid life span of a data item. But it calculates AVI only based on previous update interval. In [12], a method based on PLP (Predicted Life Period), which takes care of the

dynamicity of the life time of data was proposed. Here, life span of data is predicted based on the probability of updation of data item. This method makes PLP of data item very close to the actual valid life span of a data item.

In [10], a transaction model for supporting mobile collaborative works was proposed. This model makes use of Export-Import repository, which is a mobile sharing work space for sharing data states and data status. But in the Export-Import repository based model, locking is the main technique which has the following disadvantages. (i) More bandwidth is needed for request and reply since the locking and unlocking requests have to be sent to the server. (ii) Disconnection of mobile host or a transaction failure will result in blocking of other transactions for a long period. Our framework is better than the model which uses Export-Import repository for sharing data since it minimizes message communication costs and data update costs to a larger extent. Also disconnected mobile hosts are treated separately within a mobile affiliation by waiting for their reconnection. Transaction Management solution proposed in [15], is for reducing energy consumption at each MHs by allowing each MH to operate in three modes, Active, Doze, and Sleep thus providing a balance of energy consumption among MHs.

III. MOBILE AGENT BASED ARCHITECTURE

The proposed Mobile Agent based architecture model is illustrated in Fig 1. The model consists of Mobile Hosts,

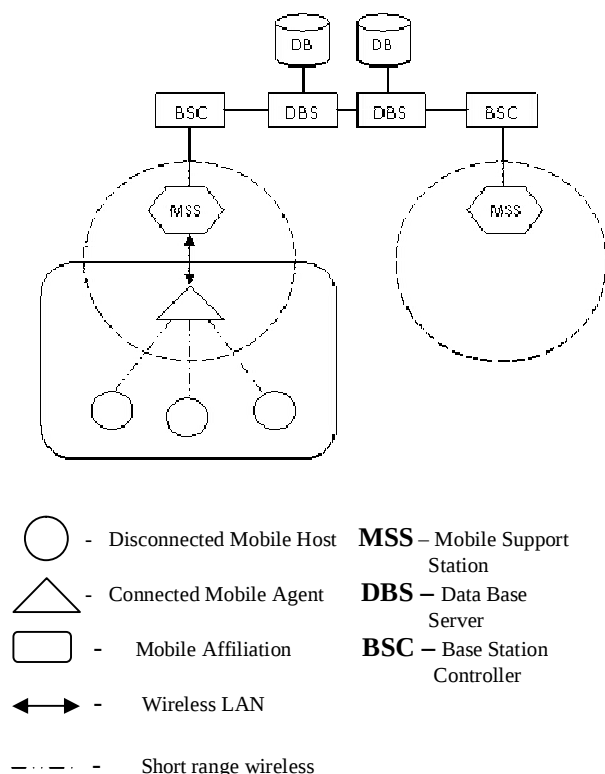


Fig.1. Mobile Agent Based Architecture model

Mobile Support Stations (MSS), and Data Base Servers (DBS). MSS is connected with a Base Station Controller (BSC) [20], which coordinates the operations of MSS using its own stored program. Unrestricted mobility is supported by wireless link between MSS and Mobile Hosts. Each MSS serves one cell whose size depends on the power of its MSS. Data Base Servers are connected to the mobile system through wired lines as separate nodes. Each DBS can be reached by any MSS and new DBSs can be connected and old ones can be taken out from the network without affecting mobile communication. A DBS can communicate with a MH only via MSSs.

Mobile agent is a special mobile node which connects to the MSS to cache the frequently accessed data. Disconnected Mobile Hosts can connect to the Mobile Agent using short range wireless communication technologies to form mobile affiliation workgroup.

Mobile hosts are allowed to access data from the cache. When data request is made for the first time, data is retrieved from the server and stored in the cache. Subsequent requests are handled by the Data Access Manager module itself. When a mobile host requests for data update, after local updation of the data item, invalidation report is sent to all the mobile hosts that have already accessed the same data. This makes all the mobile hosts to refresh their data values. When a mobile host is disconnected from the Mobile agent after updation request, the updation task is transferred to the Data Access Manager in the Mobile Agent. Data Access Manager module is used to coordinate the operations in the cache.

After disconnected from the server, Mobile Agent can move along with the connected MHs and MHs can continue their transaction execution. If data update at the server is requested, mobile agent will wait for reconnection before updation is made.

IV. MOBILE TRANSACTION FRAMEWORK

When Mobile Hosts enter into the Mobile Agent area, they connect to the Mobile Agent using short range wireless network technology to form Mobile Affiliation Work Group. Frequently accessed data are cached in the Mobile Agent. Mobile Hosts can access the cached data in the Mobile Agent. The Data Access Manager module at the Mobile Agent is responsible for enforcing concurrency and cache invalidation.

A. Energy and Connectivity Evaluation

Mobile Hosts all the time maintains its energy availability and connectivity. Connectivity is evaluated based on signal strength. When signal strength goes below one fourth of total strength, connectivity is considered as Low. When available energy goes below 25% of total energy level, then energy availability is considered as Low. The status of an MH based on Energy Availability and Connectivity (A_{ij}) can be A_{00} – Low Energy & Low Connectivity, A_{01} – Low Energy & High Connectivity, A_{10} – High Energy & Low Connectivity and A_{11} – High Energy & High Connectivity. When Data Access Manager receives a transaction request from a mobile host, it assigns a priority level using A_{ij} . A mobile host with low

energy and low connectivity is assigned the highest priority. Other levels of priority are assigned according to the various possibilities as given in Table I.

TABLE I
LEVELS OF PRIORITY

Status of an MH(A_{ij})	Energy Availability(i)	Connectivity (j)	Priority
A_{00}	Low	Low	1
A_{01}	Low	High	2
A_{10}	High	Low	3
A_{11}	High	High	4

B. Concurrency Control Mechanism

When more number of mobile hosts are accessing data simultaneously the problem of data inconsistency arises. This problem can be solved if we use an efficient concurrency control mechanism. When data request is made for the first time, data is retrieved from the server and stored in the cache. Future requests for data are managed directly by the Data Access Manager.

Data Access Manager uses a suitable data item format to store data as quintuple [12] in the cache. It has (id, TLU, PLP, dataval, NT) where id denotes unique Id of the data item, TLU indicates time of Last Update, PLP is Predicted Life Period, dataval is current value of the data item and NT denotes number of transactions that concurrently access the data item.

When Data Access Manager fetches data for the first time from the server, it sets TLU to current time, PLP to optimal time based on the nature of data item and NT to 1. NT is incremented whenever a new data access request is made. Data in the cache becomes invalid, once it is updated in the server. Life span of a data item is predicted using PLP. It makes use of the probability of updation as a basis for setting valid life span of a data item. In PLP interval, data item is valid and all the mobile hosts can access same data item concurrently.

When a MH makes update request or PLP expires, the data item is invalidated. Now PLP is modified and invalidation report is sent. The predicted life period of data item is computed using the following formula given in [12].

$$PLP = PPLP \pm (p * PPLP)$$

Where PPLP is Previous Predicted Life Period and p is predicted probability of updation of data item. $p = \text{Total_updates} / \text{NT}$. It is the ratio of data item update to data item access. Since predicted probability of updation is based on recent past history of updation rate, it is highly probable that PLP is very close to the actual validity interval of the data item.

C. Transaction Execution in the MH

After connecting to the Mobile Agent, the MH intimates the status of Energy availability and Connectivity (A_{ij}) of the MH to DAM at Mobile Agent. Then, the execution of the

transaction starts locally. If the operation is Data Read, the request is sent to DAM at the Mobile Agent. Otherwise, If the MH wants to update data, it first checks whether it is a transaction update or not. If it is a transaction update, it will check if the MH is about to be disconnected. If so, the updation task is assigned to the Data Access Manager before disconnection. Otherwise, update request is sent to Data Access Manager. The algorithm is given in Fig. 2.

```
Trans_Req_from_MH_to_MA(T, MH_ID)  
// Transaction T is initiated by an MH whose ID is MH_ID //  
Begin  
  Connect to the Mobile Agent (MA')  
  Intimate the status of Energy availability and Connectivity (Aij)  
  of the MH to DAM at MA'  
  Start execution of the transaction T locally  
  If Data Read  
    Submit Read Request to DAM at MA'  
  Else If Data Update  
    If MH is about to be disconnected  
      Assign updation task to DAM at MA' and disconnect  
    Else  
      Submit Update Request to DAM at MA'  
    End If  
  Else if commit  
    Commit  
  Else  
    Exit  
  End if  
End if  
End
```

Fig.2. MH Execution algorithm

D. Function of Data Access Manager and Server

After receiving a transaction request from an MH, the transaction is scheduled by assigning a priority to it by checking the energy availability and connectivity (*A_{ij}*). Then, the transaction is placed in a Priority queue. After scheduling, DAM will take the first transaction from the queue. When MH makes a read request, if it is in the cache of the Mobile Agent, NT is incremented by one. Otherwise, if Mobile Agent is connected to the server, it will fetch the data from the server and initialize data item format. If Mobile Agent is disconnected, the read request will be put in the queue and it will wait for Mobile Agent to get reconnection with the server. Once reconnection is got, it will fetch data from the server and initialize data item format.

When MH makes an update request, DAM updates data locally and invalidation report is sent to all the mobile hosts that have already accessed the same data item. This forces all the transactions to refresh their data values. In order to forward this update request to the server, it will check whether Mobile Agent is connected to the server. If so, update request is forwarded to the server. Otherwise, the update request will be put in the queue and it will wait until reconnection of

Mobile Agent with the server is established. After reconnection with the server, update request is forwarded to the server. The server updates the data and sends invalidation confirmation along with the updated value. Once Data Access Manager receives the confirmation, it updates the data in the cache. The data in the cache is invalidated if updation is made in the server or PLP expires.

If transaction update is made by the Data Access Manager for the disconnected MH, the above procedure is followed except that at the end, DAM generates updation report and forwards it to the MH when it gets reconnected. The algorithms are shown in Fig. 3 and 4.

```
Data_Access_Manager(T, MH_ID, Aij)  
// DAM module will be executed when MA receives a request from  
transaction T with Requesting Mobile Host ID MH_ID, Status of  
Energy availability and Connectivity Aij //  
Begin  
  Schedule the transaction T by assigning priority to it by checking  
  Aij and place it in a priority queue  
  Take the first transaction T1 from the queue  
  If Read Request  
    If data is in the cache of MA  
      Update NT  
    Else  
      If MA is disconnected  
        Wait for reconnection  
        If reconnected with server  
          Fetch data from server and initialize quintuple  
        End if  
      End if  
      Send data to MH  
    End if  
  Else If Connected or disconnected Update request  
    Update locally and send invalidation report  
    If MA is disconnected  
      While reconnection with the server is not got do  
        Wait for reconnection  
      End while  
    End if  
    Forward update request to server  
    If Server Update / PLP expiration  
      Update quintuple  
    Else  
      Wait for MH request  
    End if  
    Check for any disconnected updates by MHs  
    if disconnected updates  
      Generate and forward update reports to MHs  
      when reconnection  
    End if  
  Else  
    Wait for MH request  
  End if  
End if  
End
```

Fig.3. Data Access Manager Algorithm

Server_Execution()

Begin

Wait for connection

If connection request

Connect authorized Mobile Hosts

If Data Read

Send data item

Else if Data Update

Update data and send invalidation confirmation
with updated data

End if

End if

End if

End

Fig.4. Server execution algorithm

V. PERFORMANCE ANALYSIS

Simulation for the framework is done in Pentium Dual Core System @ 2.4 GHz with 3 GB RAM using J2ME and NS2. A mobile network is simulated with 25 nodes and 4 Mobile Agents. Mobile nodes move using random walk mobility model [21]. Response time is calculated as the time taken to service the request made by the mobile host. The response time is evaluated for the executed transactions for the E-I repository model and the proposed scheme for both non disconnected and disconnected cases. The results of the analysis for the two cases are shown in Fig 4 and 5.

VI. DISCUSSION

Fig. 4 shows the comparison of the performance of E-I repository model [10] with the proposed model for non disconnected case. As the number of transactions increases, the response time increases steadily for both E-I model and the proposed model. But the proposed model suffers from a slight increase in response time compared to the E-I model, when the number of transactions exceeds 11. This is due to the presence of Agent delay. It is also found that E-I model takes more response time than the proposed model up to 10 transactions. This is due to the extra time involved in communication overhead for locking mechanism of E-I repository model. The cache in the Mobile Agent also contributes to the low response time taken by the proposed model.

Fig. 5 for transactions with disconnection case shows same trend as non disconnected case discussed above. But it is found that the disconnection support is provided by the proposed model only up to 12 transactions. And there is a steep increase in response time, when the number of transactions exceeds 12. This is due to the communication overhead associated with more number of updation tasks that are transferred to the Data Access Manager in the Mobile Agent during disconnection, as the number of transactions increases.

The response time for Mobile Hosts with each priority level as discussed above are evaluated for 25 cases. The average response time is estimated for each priority level is

given in table II. It is found that MHs with priority level 1 get the least average response time while MHs with priority level 4 have got the highest average response time. The average response time for the MHs increases with the increase in priority level. This is due to the fact that scheduler gives more emphasis to top priority requests. Hence the proposed model is suitable for mobile networks with poor energy and poor connectivity MHs.

TABLE II
AVERAGE RESPONSE TIME FOR MOBILE HOSTS WITH PRIORITIES

Priority level	1	2	3	4
Avg. Response time in Sec.	4.5	7.6	10.8	15

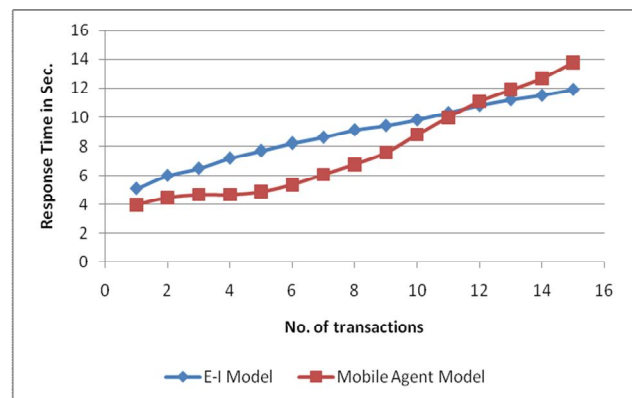


Fig.4. Analysis of Response time for Transactions without disconnection

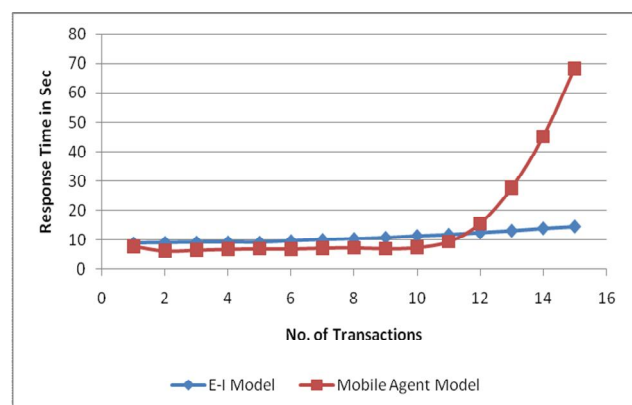


Fig.5. Analysis of Response time for Transactions with Disconnection

VII. CONCLUSION

In this paper, we have proposed a priority based transaction scheme for disconnected mobile computing environment. Mobile Agent can form a Mobile Affiliation work group with the disconnected Mobile Hosts using short range wireless technology. The frequently accessed data are cached in the Mobile Agent. This cached data can be accessed by the mobile hosts when they get connected. Since the proposed scheme uses priority based scheduling for transactions, we get better results for mobile networks in which MHs have poor energy and poor connectivity. When mobile hosts are disconnected from the Mobile Agent, transaction update task can be transferred to the Mobile Agent. By using a Mobile Agent and concurrency control without locking for accessing data, we claim that message communication costs and database update costs are minimized to a larger extent.

REFERENCES

- [1] Ho-Jin Choi, Byeong-Soo Jeong, "A Timestamped Based Optimistic Concurrency Control for Handling Mobile Transactions," ICCSA 2006, LNCS 3981, pp. 796-805, 2006.
- [2] Serrano-Alvarado, P., C. Roncancio, and M. E. Adiba, "A Survey of Mobile Transactions," *Distributed and Parallel Databases*, Vol. 16, No. 2, pp. 193-230, 2004.
- [3] Sanjay Kumar Madria, Mukesh Mohania, Sourav S. Bhowmick, Bharath Bhargava, "Mobile data and transaction management," Elsevier Information Sciences [4], pp. 279-309, 2002.
- [4] Dunham, M. H., and V. Kumar, "Impact of Mobility on Transaction Management," *MobiDE*, pp. 14-21, 1999.
- [5] Vijay Kumar, *Mobile Database Systems*, WILEY INTERSCIENCE, 2006.
- [6] Victor C.S., Kwok wa Lam and Son, S.H., "Concurrency Control Using Timestamp Ordering in Broadcast Environments," *The Computer Journal*, Vol.45 No.4, pp. 410-422, 2002.
- [7] P.A Bernstein, V. Hadzilacos and N. Goodman, *Concurrency control and Recovery in Database Systems*, Addison Wesley, 1987.
- [8] Chrysanthi P.K., "Transaction processing in a mobile computing environment," in *Proc. IEEE Workshop on Advances in Parallel and Distributed Systems*, October 1993, pp. 77-82.
- [9] Madria, S. K., M. Baseer, and S. S. Bhowmick, "A Multiversion Transaction Model to Improve Data Availability in Mobile Computing," *CoopIS/DOA/ODBASE*, pp. 322-338, 2002.
- [10] Le, H. N., and M. Nygård, "A transaction model for Supporting mobile Collaborative Works," *IEEE*, pp. 347-355, 2007.
- [11] Salman Abdul Moiz, Mohammed Khaja Nizamuddin, "Concurrency Control without Locking in Mobile Environments," *IEEE*, pp. 1336-1339, 2008.
- [12] Miraclin Joyce Pamila J.C and Thanuskodi K, "Framework for transaction management in mobile computing environment," *ICGST-CNIR Journal*, pp. 19-24, 2009.
- [13] V. Kumar, *Performance of concurrency control mechanisms in Database Systems*, Prentice Hall, Englewood Cliff, NJ, 1996.
- [14] H.T. Kung and John T. Robinson, "On optimistic methods of concurrency control," *ACM Transaction of Database Systems*, Vol. 6, No.2, 1981.
- [15] Le Gruenwald, Shankar M. Banik, Chuo N. Lau, "Managing real time database transactions in mobile ad-hoc networks," *Springer*, 27-54, 2007.
- [16] D.J. Rosencrantz, R.E. Sterns, and P.M. Lewis, "System level concurrency control for distributed database systems," *ACM Transaction of Database Systems*, Vol. 3, No. 2, 1978.
- [17] P.A. Alsberg and J.D. Day, "A principle for resilient sharing of distributed resources," in *Proc. 2nd International Conference on Software Engineering*, 1976.
- [18] A Borr, "High Performance SQL through Low Level System Integration," in *Proc. ACM SIGMOD Conference*, June 1988.
- [19] T. Ozsu and Valduriez, *Principles of Distributed Database Systems*, Prentice Hall, Englewood Cliffs, NJ, 1999.
- [20] Rajan Kurupillai, Mahi Dontamsetti and Fil J. Cosentino, *Wireless PCS*, McGraw-Hill, New York, 1997.
- [21] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad-hoc network research," *Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, vol. 2, no. 5, pp. 483-502, 2002.

DESIGN OF CONTENT ORIENTED INFORMATION RETRIEVAL BASED ON SEMANTIC ANALYSIS

S.Amudaria,
PG Student,
Dept of IT, SSN College of Engineering,
Chennai, India
daria.amu@gmail.com

S.Sasirekha,
Asst. Professor,
Dept of IT, SSN College of Engineering,
Chennai, India
sasirekhas@ssn.edu.in

Abstract:

The existing Information Retrieval (IR) systems which are based entirely on syntactic (keyword based) contents have serious limitations like irrelevant document retrieval, word sense ambiguity, low precision and recall ratio since the complete semantics of the contents are not represented. To overcome these limitations, from the recent literature it is identified that it is necessary to analyze and determine the semantic features of both the content in document and query. Hence in this paper it is proposed to initially develop a semantic pattern that represents semantic features of the contents in every document in the corpus as a Term Document Matrix (TDM) format. Then to develop a semantic pattern for the contents in the query by incorporating it with Natural Language Processing technique along with Synset (WordNet) for query refinement & expansion. Now the similarity between the semantic pattern of the query and TDM is calculated using Latent Semantic Analysis (LSA) and plotted in Semantic Vector Space. Then by matching against the vector space, contents associated to the query can be identified in the corresponding cluster. Various experimental results are carried on, which shows the increase in document retrieval recall and precision rates, thereby demonstrating the effectiveness of the model.

Keywords: Information retrieval, Semantic extraction, Query extension, Query matching

I INTRODUCTION

The existing information retrieval systems are mostly keyword-based and identify relevant documents or information by matching keywords. Keyword-based search, in spite of its merits of expedient query for information and ease-of-use, has failed to represent the complete semantics contained in the content (Oh et al, 2007) and has led to the following problems (Abdelali et al, 2007; Moreale et al, 2004): (1) keywords could represent only fragmented meanings of the content, and the content identified through keywords did not always meet the querist requirements. The querist had to screen retrieval results and correct keywords several times to obtain the required information. (2) Compared to a text, a query usually comprised fewer contents, which might lead to wrong retrieval results due to problems like insufficient information being used in the search process, insufficient query topics, and difficulty in determining query features. (3) Due to synonym and polysemy in human language, information retrieval through keywords can only cover information containing the same keyword, while other information with similar semantics but

different keywords has been completely left out. The user normally goes to the search engine to get the exact and relevant results. But the current search engine is not responsible for producing the accurate results to the user.

Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results. Rather than using ranking algorithms such as Google's Page Rank to predict relevancy, Semantic Search uses semantics, or the science of meaning in language, to produce highly relevant search results. In most cases, the goal is to deliver the information queried by a user rather than have a user sort through a list of loosely related keyword results. Here WordNet is used to get the semantics of the query.

A brief literature survey about the information retrieval techniques are done in the section II, then the proposed system is explained using various techniques in section III and the next two sections deals completely about the implementation and Test results.

II RELATED WORK

Ming-Yen Chen et al. (2009) introduces a semantic enabled information retrieval in which a web corpus is taken and the related information is retrieved. The limitation of this project is that it won't deals about the Synonyms or Synsets. Here in our project we have concentrated on WordNet ontology to collect more senses.

Zongli Jiang et al. (2009) introduce the concept of category attribute of a word. According to the category attribute of a word, the useless results can be removed from the search results and the retrieval efficiency will be improved. Latent semantic analysis is a method that can discover the underlying semantic relation between words and documents. Singular value decomposition is used in latent semantic analysis to analyze the words and documents and get the semantic relation finally.

Hongwei Yang et al. (2010) can enable the users to find the relevant documents more easily and also help users to form an understanding of the different facets of the query that have been provided for web search engine. A popular

technique for clustering is based on K-means such that the data is partitioned into K clusters. In this method, the groups are identified by a set of points that are called the cluster centers. The data points belong to the cluster whose center is closest. The algorithm used in the proposed system is K-means clustering algorithm.

Gang et al. (2009) proposed a method to enhance the information retrieval recall and precision. To filter out the document which have smaller related degree with original query, the scores of search results document is re-calculated by use of ontology semantic similarity. A new definition of the iterative query expansion parameters is put forward which can reduce the number of expansion and further improve the efficiency of the query.

Trong Hai et al. (2008) proposed a system which applies the relations between entities discovered from Text corpus to ontology integration tasks in which the noun phrase (NP) is used to identify its head noun; this is useful to avoid wrong relations between entities. It also proposes a collaborative acquisition algorithm combining WordNet-based and Text corpus to provide general concepts and their relations for ontology integration tasks.

Trong Hai Duong, Geun Sik Jo (2009) designed a new measure based on semantic ontology database WordNet is proposed, which combines information content-based measure and the edge-counting techniques to measure semantic similarity. "PART-OF" and "IS-A" hierarchical relations' influence are considered on the semantic similarity in this paper. Breadth-first search is used to find the shortest path between two concepts. The similarity of hierarchy and superposition are calculated respectively.

III PROPOSED SYSTEM

The proposed system uses the semantic analysis technique to retrieve the content which is relevant to the user query. The user's query will be analyzed in the semantic extraction and determination module to extract its semantic features for the purpose of determining contents of the query and representing them in a structured and materialized semantic pattern. In this component the semantic elements are identified and analyze their semantic relations, to be followed by the integration and simplification of semantic relations with Word Net. Now the semantic extension module will identify other potentially relevant semantic features based on semantic features of the query and include them into the query patterns. This will increase the number of semantic features in the query as the basis for matching. The input query from the user is processed using preprocessing techniques such as stop word list removal and then stemming is done. Each and every processed word is passed to the WordNet to collect all the other senses that the corresponding word has. The Synsets related to the query are taken and latent semantic Analysis process is done to index the documents.

The Singular value decomposition will process the document in the corpus and the term document frequency matrix is generated. This term document frequency matrix is plotted and most similar terms that are corresponding to the query will be plotted in the semantic space. Then finally the relevant documents are obtained by using the k-means clustering. The block diagram of the proposed system is shown below.

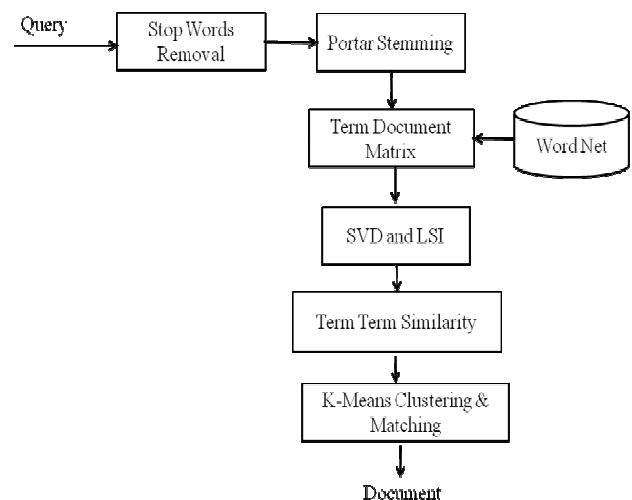


Fig1: Block diagram of the proposed system

Consider the word Java. The corresponding senses of word java that are taken from WordNet are given below.

Word	Senses
Java	an island in Indonesia south of Borneo; one of the world's most densely populated regions
Java	Coffee- a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee"
Java	a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine

It means that the single word java has three senses. This type of word ambiguity is not satisfied by the current search engines. Also consider another example. The query given by the user is Computer. Both the words PC and Computer refer to a same thing. But in the current search engines, only the documents containing the word computer will be indexed and retrieved to the user. So even though the word PC resembles the same meaning, the pages relevant to word PC are not retrieved. Hence precision and recall ratio is minimized.

The proposed system is to design a content based information retrieval based on semantic Analysis where we use WordNet ontology for performing a search based on Synsets and thus to increase the precision and recall ratio.

Precision: What fraction of the returned results is relevant to the information need?

Recall: What fraction of the relevant documents in the collection was returned by the system?

IV IMPLEMENTATION

The system is implemented by using a corpus of 250 documents. The query is given as input and the processing steps are explained below. The final output obtained is the relevant document that exactly matches with the Query.

- Semantic Pattern Construction
- Semantic Query Processing
- Semantic Query Refinement & Expansion
- Semantic Pattern Matching

1. Develop Semantic Pattern from the content

Developing a semantic pattern from the content requires the following steps. The given content is pre-processed using the porter stemming algorithm to find the root word and removing the stop words. The stop words are given manually which doesn't make any senses in the content and query.

i. Content Preprocessing

A content repository of 250 text documents is taken as corpus. These documents are to be processed in to tokens. Some selected stop words are taken. These stop words are discarded by the search engine. All the text documents that are present in the corpus are passed through these stop word list. The document word that matches with the stop word is considered to be stop word and is eliminated. This step is done to reduce the token. The remaining word is considered to be keyword and is stored in a text file. Normally the stop words will be pronouns, Articles and Prepositions.

ii. Porter Stemming Algorithm

After removing the stop words the keywords are passed to a stemming Algorithm. The stemming Algorithm used in this work is *Portar Stemming Algorithm*. This component identified semantics elements like subject, object, and predicate in the content semantics and analyzes their semantic relations.

iii. Term Document Matrix

Generate a Term Document Matrix to know the occurrences of each and every key word in the document. The term-document matrix is a large grid representing every document and content word in a collection. The TDM (Term Document Matrix) is generated by arranging the list of all content words along the vertical axis, and a similar list

of all documents along the horizontal axis. These need not be in any particular order, as long as it is kept track of which column and row corresponds to which keyword and document.

2. Query Refinement and Expansion using WordNet

i. Query Refinement

The query entered by the user is passed through the stop word list to remove the stop words. Then stemming is also done to retrieve only the subject. This is passed to the WordNet to get more senses. For example, the word vomit has 3 senses such as vomits, barf and puke. In the keyword based search only the vomit word will be taken but not its senses. Hence different words expressing the same meaning will not be taken and so the user won't be satisfied with the results of search engine. Hence pass each and every token of the query to the WordNet to get more senses.

ii. Query Vector Coordinates

The query vector coordinates are generated by checking the keyword txt file and count the occurrences of it. The senses are also counted and hence the count is incremented. The goal of WordNet project is the creation of dictionary and thesaurus, which could be used intuitively. The next purpose of WordNet is the support for automatic text analysis and artificial intelligence. WordNet is a lexical database for the English language. It groups English words into sets of synonyms called Synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every Synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as "car pool"); different senses of a word are in different Synsets.

A query Q is represented as an n -dimensional vector q in the same vector space as the document vectors. There are several ways how to search for relevant documents. Generally, we can compute matrix to represent the similarity of query and document vectors.

3. Perform SVD and LSA

i. Term Frequency – Inverse Document Frequency

After constructing the Term Document Matrix apply weight to all token found in countMatrix. The TFIDF (Term Frequency – Inverse Document Frequency) is calculated using the formula

$$TFIDF_{i,j} = (N_{i,j} / N^*,j) * \log(D / D_i)$$

Where

$N_{i,j}$ = the number of times word i appears in document j (the original cell count).

N^*,j = the number of total words in document j (just add the counts in col j).

D = the number of documents (the number of columns).

D_i = the number of documents in which word i appears (the number of non-zero columns in row i).

The TFIDF matrix obtained is used for the computation of the Singular Value Decomposition.

ii. Singular Value Decomposition

The matrix that is generated from the term frequency- inverse document frequency matrix is used for the computation of Singular value decomposition.

A rank-reduced, Singular Value Decomposition is performed on the matrix to determine patterns in the relationships between the terms and concepts contained in the text.

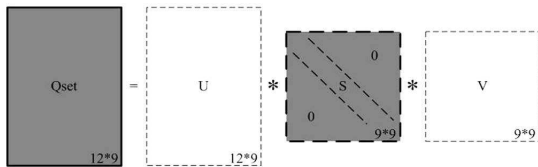


Fig.2 SVD Computation

The SVD forms the foundation for LSI. It computes the term and document vector spaces by transforming the single term-frequency matrix, A , into three other matrices— a term-concept vector matrix, T , a singular values matrix, S , and a concept-document vector matrix, D , which satisfy the following relations:

$$\begin{aligned} A &= TSD^T \\ T^T T &= D^T D = I_r \\ TT^T &= I_m \quad DD^T = I_n \\ S_{1,1} &\geq S_{2,2} \geq \dots \geq S_{r,r} > 0 \quad S_{i,j} = 0 \text{ where } i \neq j \end{aligned}$$

In the formula, A , is the supplied m by n weighted matrix of term frequencies in a collection of text where m is the number of unique terms, and n is the number of documents. T is a computed m by r matrix of term vectors where r is the rank of A —a measure of its unique dimensions $\leq \min(m,n)$. S is a computed r by r diagonal matrix of decreasing singular values, and D is a computed n by r matrix of document vectors.

The LSI modification to a standard SVD is to reduce the rank or truncate the singular value matrix S to size $k \ll r$, typically on the order of a k in the range of 100 to 300 dimensions, effectively reducing the term and document vector matrix sizes to m by k and n by k respectively. The SVD operation, along with this reduction, has the effect of preserving the most important semantic information in the text while reducing noise and other undesirable artifacts of

the original space of A . This reduced set of matrices is often denoted with a modified formula such as:

$$A \approx A_k = T_k S_k D_k^T$$

Efficient LSI algorithms only compute the first k singular values and term and document vectors as opposed to computing a full SVD and then truncating it.

iii. Latent Semantic Analysis (LSA)

By reducing the term-document space to fewer dimensions, SVD reveals the underlying relationships between terms and documents in all possible combinations and the similarity between terms and documents are shown within the reduced space. This technique uses a term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to terms and whose columns correspond to documents. Latent semantic analysis (LSA) is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

A typical example of the weighting of the elements of the matrix is tf-idf (term frequency–inverse document frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up weighted to reflect their relative importance. The inverse weighted term document matrix calculates the occurrence of single word in all the documents.

4. Query Projection and Matching

In the LSI model, queries are formed into *pseudo-documents* that specify the location of the query in the reduced term-document space. Given q , a vector whose non-zero elements contain the weighted term-frequency counts of the terms that appear in the query, the pseudo-document can be represented by

$$A = TSD^T$$

The singular values are used to individually weight each dimension of the term-document space. Once the query is projected into the term-document space, one of several similarity measures can be applied to compare the position of the pseudo-document to the positions of the terms or documents in the reduced term-document space.

i. Term-Term Similarity

After getting U and S matrix, multiply U and S matrix with the resultant matrix (say T matrix) to find Term-Term Similarity using Cosine relation.

$$\cos A = \frac{V \cdot W}{\|V\| \|W\|}$$

If the angle is between 0 and 90 then there exists a relation (some similarity) between the two vectors coordinated. Similarly cosine relation is computed between query coordinate and other coordinate. Lesser the angle more similarity between the terms. The cosine similarity measure, is often used because, by only finding the angle between the pseudo-document and the terms or documents in the reduced space, the lengths of the documents, which can affect the distance between the pseudo-document and the documents in the space, are normalized. Once the similarities between the pseudo-document and all the terms and documents in the space have been computed, the terms or documents are ranked according to the results of the similarity measure, and the highest-ranking terms or documents, or all the terms and documents exceeding some threshold value, are returned to the user

ii. K-means Clustering

The vector coordinates whose cosine similarity value is greater than the threshold are retrieved and plotted in the semantic space to make the search to be more relevant. Then the k-means clustering is done to make the cluster documents much relevant to the query.

The basic idea of k-means algorithm is to do a local optimization on a given number of clusters. Specifically, first randomly pick up k documents from the entire collection and make them as the initial centroid of the desired k clusters. Then for each document in the collection find the nearest centroid and put this document into the corresponding cluster. After each document is assigned to one of the cluster, recompute the centroid and repeat the computation. This method iteratively optimize the clusters until the computation converge when the clusters do not change anymore and the clustering quality achieved a local maximum. The advantage of the k-means is that its complexity is very low and is very easy to implementation. In a cluster, the similarity between the query and each content in the cluster is computed to sort contents by the order of similarity and offer the most approximate content to the querist.

V TEST RESULTS

The proposed system is implemented by using a corpus of 250 documents. The tokens are separated from the corpus by means of Keywords. The screenshots are shown below.

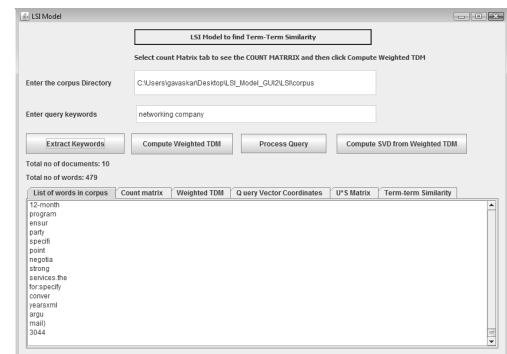


Fig.3: Keywords Extraction

The keywords are collected and Term document matrix is generated. Finally the relevant document is retrieved. This also handles the Synonym by passing the query through the WordNet.

Precision and Recall ratio

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced by a web search engine for a query) and a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic). In the context of information retrieval, *precision* is defined as the ratio of relevant documents to the number of retrieved documents:

$$\text{Precision} = \frac{\text{Number of relevant documents}}{\text{Number of retrieved documents}}$$

and *recall* is defined as the proportion of relevant documents that are retrieved:

$$\text{Recall} = \frac{\text{Number of relevant, retrieved documents}}{\text{Total number of relevant documents}}$$

Consider 250 documents and the relevant document retrieved is 70. Here the precision and recall ratio are calculated as 0.28 and 0.7 respectively. Hence the average precision and recall ratio is 0.378 and is depicted in the below graph.

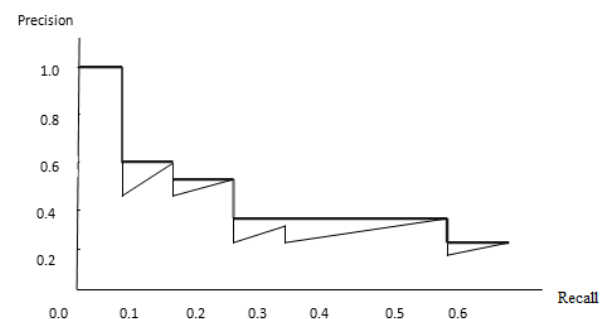


Fig 4: precision-recall ratio curve

VI CONCLUSION

In this study, the proposed approaches can efficiently and precisely perform semantic based information retrieval. In addition to semantic-based information retrieval, the proposed system has two significant parts: a semantic extension model which employs latent semantic analysis to generate more semantics for matching, thereby solving the problem of insufficient information for query; and a semantic clustering model which uses k-means clustering algorithm based on neighbours and then performs content matching in that category, thereby improving matching accuracy. Since the query is passed through WordNet all the senses will be taken and accuracy of the relevant pages will increase.

REFERENCES

- [1] Ming-Yen Chen, Hui-Chuan Chu, Yuh-Min Chen (2009), "Developing a semantic-enable information retrieval mechanism", Elsevier Journal on Expert Systems with Applications, May 2009.
- [2] Zongli Jiang and Changdong Lu, "A latent semantic analysis based method of getting the category attribute of words" 2009 International Conference on Electronic Computer Technology.
- [3] Hongwei Yang, "A document clustering algorithm for web search engine retrieval system", 2010 International Conference on e-Education, e-Business, e-Management and e-Learning.
- [4] Jianpei Zhang; Zhongwei Li; Jing Yang; "A divisional incremental training algorithm of support vector machine" Mechatronics and Automation, 2005 IEEE International Conference.
- [5] Gang Lv 1, Cheng Zheng 2, Li Zhang 3, "Text information retrieval based on concept semantic similarity" 2009 Fifth International Conference on Semantics, Knowledge and Grid.
- [6] Trong Hai Duong, Geun Sik Jo, Ngoc Thanh Nguyen, "A Method for Integration across Text Corpus and WordNet-based Ontologies" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
- [7] Zhongcheng Zhao "Measuring Semantic Similarity Based On WordNet" 2009 Sixth Web Information Systems and Applications Conference
- [8] Trong Hai Duong, Geun Sik Jo, "Semantic similarity methods in WordNet and their application to information retrieval on the web" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
- [9] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang, Shou-Bin Bong; "Toward a Semantic Search Engine based on Ontologies" Network Engineering and Research Center, South China University of Technology, Guangzhou 510640, China
- [10] Qinglin Guo, Ming Zhang (2007), "Multi-documents Automatic Abstracting based on text clustering and semantic analysis", Elsevier Journal on Knowledge Based Systems, 22, 482-485.
- [11] Berry, M.W.(1992), "Large scale singular value computations, " .International Journal of Supercomputer Applications, 6 (1), pp 13-49
- [12] Jiuling Zhang, Beixing Deng, Xing Li "Concept Based Query Expansion Using WordNet" 2009 International e-Conference on Advanced Science and Technology
- [13] Abdelali, A., Cowie, J., & Soliman, H. S. (2007). Improving query precision using semantic expansion. Information Processing and Management, 43, 705-716

Enhanced Load Balanced AODV Routing Protocol

Iftikhar Ahmad and Humaira Jabeen
Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University Islamabad, Pakistan
ify_ia@yahoo.com, humaira_jabeen_83@yahoo.com

Abstract— Excessive load on the MANET is the main reason for link breakage and performance degradation. A congested node in the network dies more quickly than other nodes. A good load balancing technique share the traffic load evenly among all the nodes those can take part in transmission. Transferring of load from congested nodes to less busy nodes and involvement of other nodes in transmission that can take part in route can improve the overall network life. We proposed a load balancing scheme for AODV that improves overall network life, throughput and reduce average end to end delay.

Keywords: AODV, load balancing, congestion, delay.

I. INTRODUCTION

A mobile ad hoc network is defined as a collection of mobile nodes with no central management, running on batteries and changing topology [12]. The routing in MANET is difficult due to its changing topology. There are three types of protocols for MANET proactive, reactive and hybrid routing protocols. Proactive routing protocols stores and maintains routing state of every other node in the network. Reactive routing protocols discover route on demand, when there is need. Hybrid routing protocols combines advantages of both reactive and proactive classes of protocols. In this study we proposed a load balanced ad hoc routing protocol by modifying basic AODV routing protocol. AODV [13] is the main reactive routing protocol for mobile ad hoc networks which is most widely used for routing in MANET. It is especially designed for MANET and performs well then other routing protocols for MANET.

The basic function of AODV is depended on two mechanisms; one is route discovery mechanism and second is route maintenance mechanism. Both of these mechanisms works through four different type of messages those are RREQ, RREP, Route error message and Hello message. Whenever a node wants to transmit data to any other node in the network, it starts route discovery process by sending a broad cost of RREQ to all its neighbors those are within transmission rang. A route reply is sent back to the source node by the destination or any intermediate node that have fresher route to destination. The reply is sent through the route which is having less number of hops.

In this way a route with less number of hops is selected during the route discovery mechanism. In the route discovery mechanism the route is discovered and selected through route discovery algorithm. Lot of work has been done on this algorithm and improvements are made in order to increase performance of protocol. Certain Load balancing schemes

are proposed to achieve good load balancing in MANET. In Load balancing we transfer the jobs from overloaded nodes to less busy nodes or idle nodes. In the result, total time to process all jobs can be reduced and also guarantee that no node will remain idle while some jobs are there to process.

Numbers of algorithms are proposed for load balancing that consider traffic load for route selection, but these algorithms are not suitable for large scale transmissions. While selecting the route we must consider that the distribution of load should be even. Mobile nodes having low traffic load should be preferred to the heavily loaded mobile nodes.

II. RELATEDWORK

Dynamic Load-Aware Routing [2] protocol, DLAR defined the network load of a mobile node as the number of packets in its interface queue. Load-Balanced Ad hoc Routing protocol [1] LBAR defined network load in a node as the total number of routes passing through the node and its neighbors.

In Load-Sensitive Routing protocol [3] the network load in a node is defined as the summation of the number of packets being queued in the interface of the mobile host and its neighboring hosts. Even though the load metric of LSR is more accurate than those of DLAR or LBAR, but it does not consider the effect of access contentions in the MAC layer. Therefore, LSR produce contention delay. WLAR [4] distributes traffic among mobile nodes through load balancing mechanism which is product of average queue size and number of shared nodes. Load Aware Routing in Ad hoc (LARA) networks protocol [5] defines a new metric called traffic density that is used to select the route with minimum load. Traffic density means the degree of contention at the medium access control layer.

Simple Load-balancing Approach (SLA)[6] not allowing traffic to be concentrated on the node and allowing each node to drop RREQ or to give up packet forwarding depending on its own traffic load to save energy. It also suggests a payment scheme called Protocol-Independent Fairness Algorithm (PIFA) for packet forwarding.

A novel load-balancing technique [7] for ad hoc on-demand routing is very effective method to achieve load balance and congestion alleviation. If a node ignores RREQ messages within a specific period, it can completely be excluded from the additional communications that might have occurred for that period otherwise. A node can decide not to serve a traffic flow by dropping the RREQ for that flow. The interface queue occupancy and workload on node is used to control RREQ messages adaptively.

Delay-based load-aware on-demand routing (D-LAOR) protocol [8], discovers the optimal path based on the estimated total path delay and the hop count.

AOMDV routing protocol [9] used Queue Length and Hop Count value together to select a route from source to destination that avoids congestion and load balancing. A threshold value is defined after a threshold alternate path is chosen. Intermediate nodes avoids broad cast of RREQ if the routes are already congested.

Aggregate Interface Queue Length (AIQL) scheme has been proposed in this paper [10] to deal with load balancing issues. A route is selected on the bases of AIQL to transmit the data. AIQL is the sum of queue length of all nodes in the path from source to destination.

All proposed protocols work well for small scale transmissions but in case of large scale transmission the ad-hoc on demand distance vector load and mobility (AODVLM) routing protocol [11] shows better results in terms of throughput and delivery ratio with little increase in routing overhead. The proposed load balancing scheme in this paper further extends the AODVLM implementation.

III. PROPOSED LOAD BALANCING SCHEME

The proposed modification extends AODV and distributes the traffic among ad hoc nodes through a simple load balancing mechanism. The protocol adopts basic AODV procedure.

A. Selecting Route Selection Procedure

When a source node initiates a route discovery procedure by flooding RREQ messages, each node that receives the RREQ looks in its routing table to see if it has a fresh route to the destination.

If it doesn't have the route it calculates the number of packets in its interface queue and divides it with its queue length and adds calculated ratio in RREQ and broadcasts it further. The process is repeated till either the destination is reached or no destination is found.

B. Averaged Aggregated Load Ratio (AALR)

If P are packets in the queue of a node and L is the length of queue then ratio of the load on node is $R = P/L$ and sum of ratio on each node in the route is $= \sum R$ then

$AALR = \sum R/N$, where N is number of hops the RREQ has passed through.

The AALR metric has been used in order to find out the heavily loaded route. Because if the aggregate queue length for the path is higher, then it obviously means that either all the nodes on the path are loaded or there is at least one node lying on the route that is overloaded.

Hence by considering a route with lesser value of averaged aggregate load ratio for selecting the path we are have diverge the packets from heavily loaded route to comparatively lighter route.

In this way traffic load is distributed among the available reachable nodes that can provide a path to destination.

Instead of increasing load on already busy nodes we are distributing traffic load among the other available nodes.

C. Use of less Busier Nodes

During the transmission the selected route expires from time to time to check the availability of less busier node for further transmission of data. There is greater chance that more nodes come closer to the active route that can provide better route for transmission. For this purpose we expires route after fixed intervals of time during the transmission of data. Instead of using the same route for entire transmission of data new route are discovered.

The following figure 1 displaying the scenario of transmission with basic AODV routing protocol.

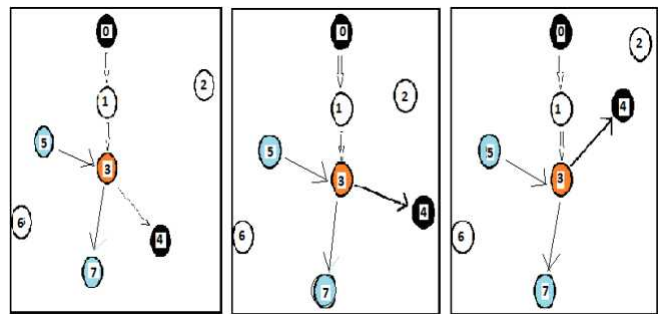


Figure 1. Example AODV Transmission Scenario.

In figure 1 node 5 is communicating with node 7 and node 0 is the source node for node 4. We can easily analyze that node 3 is busiest node that will be dead very soon. For large scale transmission there will be more bad results as far as throughput is concerned. As from the figure it is clear that node 5 can communicate through node 1 instead of node 3. With the passage of time as in part 2 of the figure the node 0 can transmit data through node 2 which is more efficient path but it does not happen in case of basic AODV. After new proposed load balancing scheme the figure 1 will be like the figure 2.

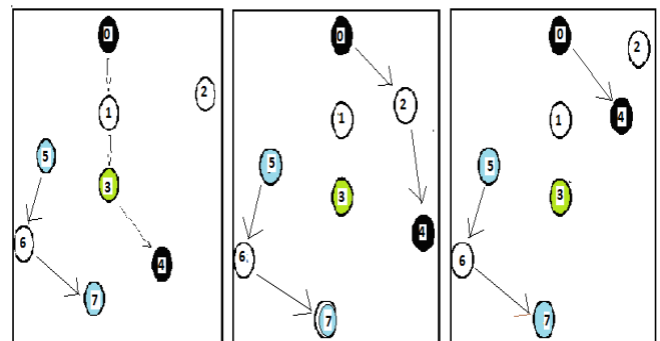


Figure 2. AODV Transmission Scenario after Implementation.

The new scheme is adopting the better and less busy route. It avoids the congested nodes during route discovery and later on during transmissions makes use of all available less busy nodes. In this way our scheme is shifting the traffic load from busier nodes to less busy nodes. New proposed scheme not using the certain nodes for entire traffic but sharing traffic load with all available nodes that can take part in communication as in shown in figure 2.

As long as a particular node remains busy means it has to transmit or forward more packets to its neighboring nodes. With transmission of each data and control packet node is consuming energy (power). Mobile node relies on batteries and battery life decrease with the forwarding of each and every packet.

So, more load means lesser life time of the node and lesser network life intern. To calculate the node's load share we calculate number of packets forward by that particular node and compare that number against each other and the network. Our proposed protocol give more even load balancing in MANET then existing load balancing protocols.

IV. PERFORMANCE EVALUATION

We implemented basic and our load balancing algorithm in NS2 [15]. NS2 is discrete event simulator for the simulation of wireless ad hoc networks. It supports Two Ray Ground propagation and Random Way Point mobility models those are required for the implementation of our work. We used the following performance metric to evaluate the performance of our load balancing algorithm against basic AODV algorithm.

A. Performance Metric

1) Average end to end delay:

This is the average overall delay occurs for a packet to travel from a source node to a destination node. This includes all possible delays caused by buffering during route discovery, queuing at the interface-queue, contention and retransmission delays at the MAC layer, and propagation and transfer times.

2) Throughput:

It is defined as the total number of packets transmitted in a given time period.

3) Traffic load Distribution:

It is the total number of packets that are forwarded by a node during transmission. Because each forwarded packet consume node's power that reduce node life.

B. Parameter Setting

The radio propagation model [14] that is considered for the protocol is the Two-Ray Ground and Random way point mobility model is used in our implementation of protocol.

The table I shows complete detail of parameter used in our simulation setting.

TABLE I. SIMULATION SETTING

Channel type	Channel/Wireless Channel
Radio-propagation model	Propagation/TwoRay Ground
Network interface	Phy/WirelessPhy
MAC type	Mac/802_11
Interface queue	Queue/Drop Tail/PriQueue
Link layer type	LL
Antenna model	Antenna/Omni Antenna
Max packet in ifq	25
Packet size	1024
Number of mobile nodes	20
Simulation time	150 seconds

C. Simulation Results

The results are compared and presented in graphical form after implementing Enhanced load balancing AODV routing protocol. We analyzed the results by taking pause time on x-axis and performance metric throughput and end to end delay on Y-axis and for load distribution number of forwarded packets on y-axis and Node ID on x-axis.

1) Throughput:

Throughput at different time interval is compared as shown in figure 5.

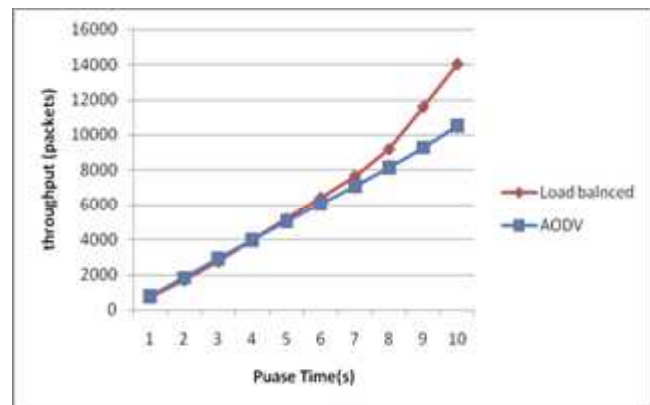


Figure 3. Throughput vs Pause time.

Total number of packets transmitted by AODV is lesser then Load balanced AODV and with passage of time throughput of load balanced AODV is increasing as more packets are transmitting in given time.

2) Average end to end delay:

Average end to end delay of load balanced is round about 18 mille seconds at the start of transmission and as transmission goes on it becomes 10 mille second but for AODV its minimum value is 14 mille seconds. It means averaged end to end delay is reduced in greater extend. This reduction in delay improves throughput that means now source node is sending packets more quickly to the destination then basic AODV.

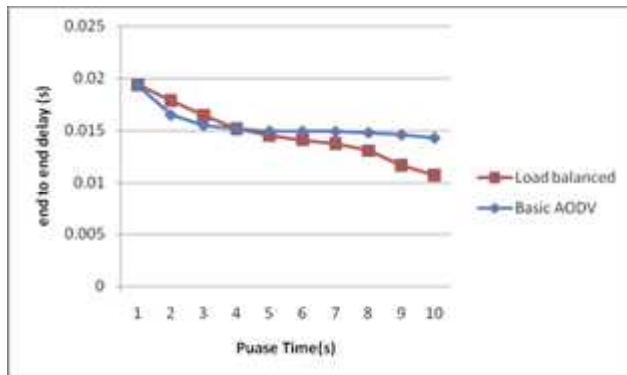


Figure 4. End to end delay vs Pause time.

3) Load distribution:

Load distribution is important because in which we analyzed how all the nodes shared the network traffic load among each other that reflects more node life. More packet forwarding means more energy consumption and more use of battery power. By not sending all the data through some specific nodes all nodes that can be involved in transmission are included in the route. Load balancing means all node sharing equal load in the network. If network is not balanced in term of traffic load some nodes have lot of load on it and some remains free. The busier nodes can get exhausted quickly and may be down quickly that result in to more link failure and performance degradation. In figure 5 traffic load is more balanced among the nodes in the network then basic AODV that show steep up and downs in the graph. In case of AODV some nodes like node 2, 5, 9 forwards only round about 100 packets and some nodes like node 3,6,8 are forwarding round about 700 packets. This means load is not balanced among the node and in case of load balanced AODV the traffic load is shared evenly among the nodes as for the all nodes number of forwarded packets are round about 400 packets.

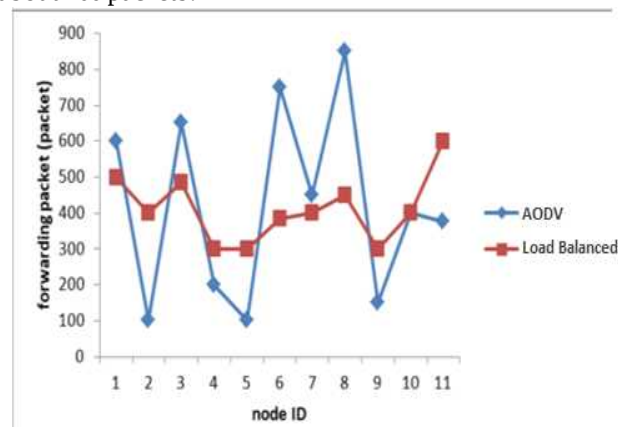


Figure 5. Forwarding packet vs Node.

V. CONCLUSION

New load balanced AODV routing protocol for distributing traffic load evenly among nodes in ad hoc networks is proposed. The idea is to provide a scheme for load distribution and to reduce congestion in high load networks. We performed a simulation study and compared the modified version of AODV with basic AODV protocol. The results of simulation shows that the proposed load balanced protocol can improve average throughput, reduce average end to end delay and improves overall network life. Hence, the proposed AODV is more useful for longer transmission and for moderately loaded high mobility networks.

REFERENCES

- [1] Hassanein, H. and A. Zhou, "Routing with Load Balancing in Wireless Ad Hoc Networks", Proceeding of ACM MSWiM, Rome, Italy, 2001 pp: 89-96.
- [2] Lee, S.J. and M. Gerla, "Dynamic Load Aware Routing in Ad Hoc Networks", Proceedings of ICC Helsinki, Finland, 2001 pp: 3206-3210.
- [3] Wu, K. and J. Hams, "Load Sensitive Routing for Mobile Ad Hoc Networks", Proceedings of IEEE ICCCN'01, Phoenix, AZ 2001, pp: 540-546.
- [4] Choi, D.I., J.W. Jung, K.Y. Kwon, D. Montgomery and H.K. Kahng, Design and Simulation Result of a Weighted Aware Routing Protocol in Mobile Ad Hoc Network: LNCS 3391, 2003 pp: 178-187.
- [5] Saigal, V., A.K. Nayak, S.K. Pradhan and R. Mall, "Load Balanced Routing in Mobile Ad hoc Networks", Elsevier Computer Communications 27, 2004 pp: 295-305.
- [6] Yoo, Y. and S. Ahn, "A Simple Load-Balancing Approach in Secure Ad Hoc Networks", ICOIN, LNCS 3090, 2004 pp: 44-53.
- [7] Lee, Y.J. and G.F. Riley, "A Workload-Based Adaptive Load-Balancing Technique for Mobile Ad Hoc Networks" Proceedings of IEEE Communication Society, WCNC 2005, pp: 202-207
- [8] Song, J.H., V. Wong and V.C.M. Leung, "Load-Aware On-Demand Routing Protocol for Mobile Ad hoc Networks", Proceedings of 57th IEEE Semiannual Vehicular Technology Conference, 2003 p3: 1753-1757
- [9] Shalini Puri, Satish R. Devene, "Congestion Avoidance and Load Balancing in AODV-Multipath Using Queue Length," Proceedings of Second International Conference on Emerging Trends in Engineering & Technology, 2009 pp. 1138-1142
- [10] Amita Rani and Mayank Dave "Performance evaluation of modified AODV for load Balancing", Journal of Computer Science 3 (11): 863-868, 2007.
- [11] Ifrikhar Ahmad and Mata ur Rehman "Efficient AODV routing based on traffic load and mobility of node in MANET", Proceedings of IEEE, International Conference on Emerging Technologies, Islamabad, Pakistan, 2010.
- [12] S. Corson and J. Macker, "Mobile Ad hoc Networking (MANET): Routing Protocol Performance issues and Evaluation Considerations", Network Working Group, RFC2501, January 1999.
- [13] C. Perkins, E. Royer and S. Das, "Ad hoc On-demand Distance Vector (AODV) Routing", IETF RFC3561, July 2003.
- [14] M. Gunes, and M. Wenig, "Models for Realistic Mobility and Radio Wave Propagation for Ad Hoc Network Simulations" Springer-Verlag London Limited, 2009.
- [15] [15] K. Fall and K. Varahan, editors. NS Notes and Documentation. The VINT Project, UC Berkeley, LBL, USC/ISI, and Xerox PARC, November 1997.

Comparative Analysis of Speaker Identification using row mean of DFT, DCT, DST and Walsh Transforms

Dr. H B Kekre

Senior Professor, Computer Department,
MPSTME, NMIMS University,
Mumbai, India
hbkekre@yahoo.com

Vaishali Kulkarni

Associate Professor, Electronics & Telecommunication,
MPSTME, NMIMS University,
Mumbai, India
Vaishalikulkarni6@yahoo.com

Abstract— In this paper we propose Speaker Identification using four different Transform Techniques. The feature vectors are the row mean of the transforms for different groupings. Experiments were performed on Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Sine Transform (DST) and Walsh Transform (WHT). All the Transform give an accuracy of more than 80% for the different groupings considered. Accuracy increases as the number of samples grouped is increased from 64 onwards. But for groupings more than 1024 the accuracy again starts decreasing. The results show that DST performs best. The maximum accuracy obtained for DST is 96% for a grouping of 1024 samples while taking the transform.

Keywords - Euclidean distance, Row mean, Speaker Identification, Speaker Recognition

I. INTRODUCTION

Human speech conveys an abundance of information, from the language and gender to the identity of the person speaking. The purpose of a speaker recognition system is thus to extract the unique characteristics of a speech signal that identify a particular speaker. [1, 2, 3] Speaker recognition systems are usually classified into two subdivisions, speaker identification and speaker verification. Speaker identification (also known as closed set identification) is a 1: N matching process where the identity of a person must be determined from a set of known speakers [3 - 5]. Speaker verification (also known as open set identification) serves to establish whether the speaker is who he claims to be [6]. Speaker recognition can be further classified into text-dependent and text-independent systems. In a text dependent system, the system knows what utterances to expect from the speaker. However, in a text-independent system, no assumptions about the text can be made, and the system must be more flexible than a text dependent system.

Speaker recognition technology has made it possible to use the speaker's voice to control access to restricted services, for example, for giving commands to computer, phone access to banking, database services, shopping or voice mail, and access to secure equipment. Speaker Recognition systems have been developed for a wide range of applications [7 - 10].

Although many new techniques have been developed, widespread deployment of applications and services is still not possible. None of these systems gives accurate and reliable results. We When you open have proposed speaker recognition using vector quantization in time domain by using LBG (Linde Buzo Gray), KFCG (Kekre's Fast Codebook Generation) and KMCG (Kekre's Median Codebook Generation) algorithms [11], [12], [13] and in transform domain using DFT, DCT and DST [14].

The concept of row mean of the transform techniques has been used for content based image retrieval (CBIR) [15 - 18]. This technique also has been applied on speaker identification by first converting the speech signal into a spectrogram [19].

For the purposes of this paper, we will be considering a speaker identification system that is text-dependent. For the identification purpose, the feature vectors are extracted by taking the row mean of the transforms (Which is a column vector). The technique is used as shown in figure 1. Here a speech signal of 15 samples is divided into 3 blocks of 5 each, and these 3 blocks form the columns of the matrix whose transform is taken. Then the mean of the absolute value of each row of the transform matrix is taken and this forms the column vector of mean.

The rest of the paper is organized as follows: Section 2 explains feature generation using the transform techniques, Section 3 deals with Feature Matching, and the results are explained in Section 4 and the conclusion in section 5.

II. TRANSFORM TECHNIQUES

A. Discrete Fourier Transform

Spectral analysis is the process of identifying component frequencies in data. For discrete data, the computational basis of spectral analysis is the discrete Fourier transform (DFT). The DFT transforms time- or space-based data into frequency-based data. The DFT allows you to efficiently estimate component frequencies in data from a discrete set of values sampled at a fixed rate. If the speech signal is represented by $y(t)$, then the DFT of the time series or samples $y_0, y_1, y_2, \dots, y_{N-1}$ is defined as:

$$Y_k = \sum_{n=0}^{N-1} y_n e^{-2\pi jkn/N}$$

(1)

Where $y_n = y_s(n\Delta t)$; $k = 0, 1, 2, \dots, N-1$.

Δt is the sampling interval.

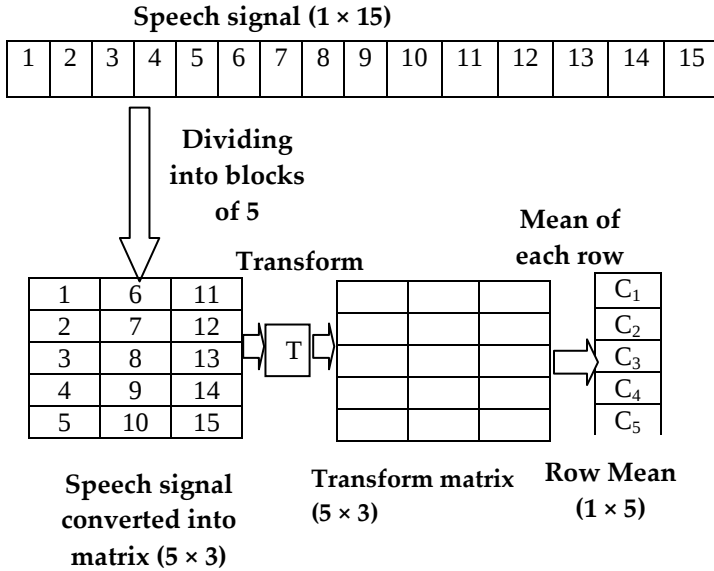


Figure 1. Row Mean Generation Technique

B. Discrete Cosine Transform

A discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies.

$$y(k) = w(k) \sum_{n=1}^N y(n) \cos \frac{\pi(2n-1)(k-1)}{2N}$$

(2)

Where $y(k)$ is the cosine transform, $k=1, \dots, N$.

$$w(k) = 1/\sqrt{N} \quad k=1$$

$$= \sqrt{\left(\frac{2}{N}\right)} \quad 2 \leq k \leq N$$

The DCT is closely related to the discrete Fourier transform. You can often reconstruct a sequence very accurately from only a few DCT coefficients, a useful property for applications requiring data reduction [20 – 22].

C. Discrete Sine Transform

A discrete sine transform (DST) expresses a sequence of finitely many data points in terms of a sum of sine functions.

$$y(k) = \sum_{n=1}^N x(n) \sin \left(\pi \frac{kn}{N+1} \right)$$

(3)

Where $y(k)$ is the sine transform, $k=1, \dots, N$.

D. Walsh Transform

The Walsh transform or Walsh–Hadamard transform is a non-sinusoidal, orthogonal transformation technique that decomposes a signal into a set of basis functions. These basis functions are Walsh functions, which are rectangular or square waves with values of +1 or −1. The Walsh–Hadamard transform returns sequency values. Sequency is a more generalized notion of frequency and is defined as one half of the average number of zero-crossings per unit time interval. Each Walsh function has a unique sequency value. You can use the returned sequency values to estimate the signal frequencies in the original signal. The Walsh–Hadamard transform is used in a number of applications, such as image processing, speech processing, filtering, and power spectrum analysis. It is very useful for reducing bandwidth storage requirements and spread-spectrum analysis. Like the FFT, the Walsh–Hadamard transform has a fast version, the fast Walsh–Hadamard transform (fwht). Compared to the FFT, the FWHT requires less storage space and is faster to calculate because it uses only real additions and subtractions, while the FFT requires complex values. The FWHT is able to represent signals with sharp discontinuities more accurately using fewer coefficients than the FFT. FWHT_h is a divide and conquer algorithm that recursively breaks down a WHT of size N into two smaller WHTs of size $N/2$. This implementation follows the recursive definition of the $2^N \times 2^N$ Hadamard matrix H_N :

$$H_N = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{N-1} & H_{N-1} \\ H_{N-1} & -H_{N-1} \end{pmatrix}$$

(4)

The $1/\sqrt{2}$ normalization factors for each stage may be grouped together or even omitted. The Sequency ordered, also known as Walsh ordered, fast Walsh–Hadamard transform, FWHT_w, is obtained by computing the FWHT_h as above, and then rearranging the outputs [23].

III. FEATURE EXTRACTION

The procedure for feature vector extraction is given below:

1. The speech signal is divided into groups of n samples. (Where n can take values: 64, 128, 256, 512, 1024, 2048, and 4096) samples.
2. These blocks are then arranged as columns of a matrix and then the different transforms given in section II are taken.

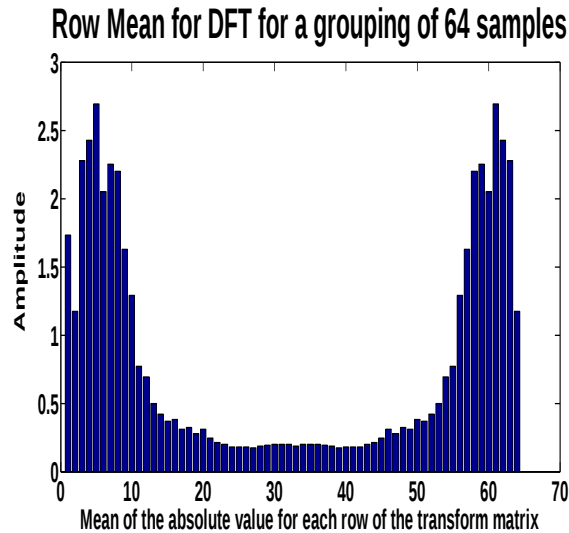
3. The mean of the absolute values of the rows of the transform matrix is then calculated.
4. These row means form a column vector ($1 \times n$ where n is the number of rows in the transform matrix).
5. This column vector forms the feature vector for the speech sample.
6. The feature vectors for all the speech samples are calculated for different values of n and stored in the database.

Figure 2 shows the row mean generated for the four transforms for a grouping of 64 samples for one of the speech signal in the databases. These 64 row means form the feature vector for the particular sample considered. In a similar fashion, the feature vectors for other speech signals were also calculated. This process was repeated for all values of n . As can be seen from figure 2, the 64 mean values form a 1×64 feature vector.

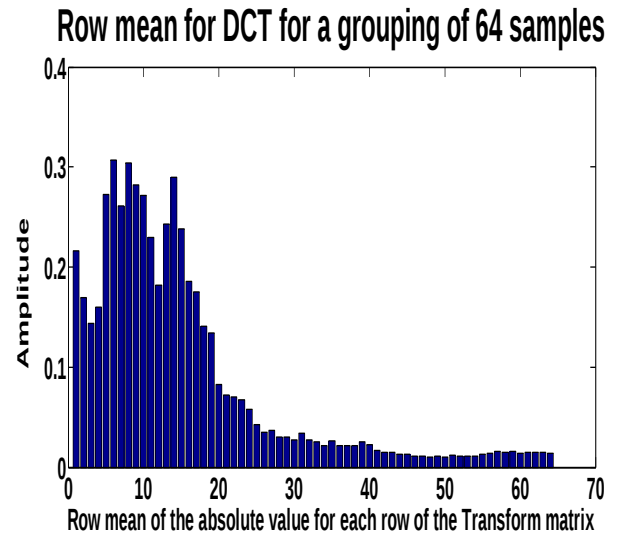
IV. RESULTS

A. Basics of speech signal

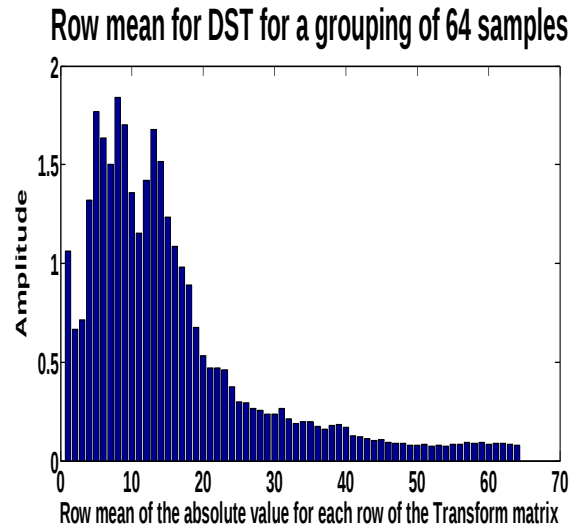
The speech samples used in this work are recorded using Sound Forge 4.5. The sampling frequency is 8000 Hz (8 bit, mono PCM samples). Table I shows the database description. The samples are collected from different speakers. Samples are taken from each speaker in two sessions so that training model and testing data can be created. Twelve samples per speaker are taken. The samples recorded in one session are kept in database and the samples recorded in second session are used for testing.



(A)



(B)



(C)

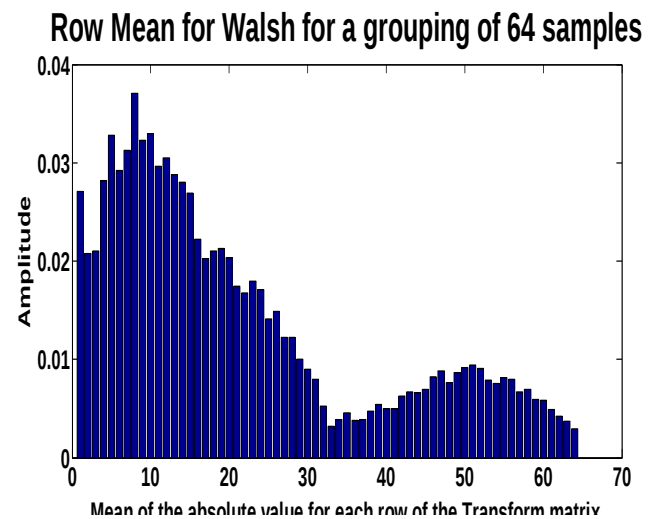


TABLE I. DATABASE DESCRIPTION

Parameter	Sample characteristics
Language	English
No. of Speakers	105
Speech type	Read speech
Recording conditions	Normal. (A silent room)
Sampling frequency	8000 Hz
Resolution	8 bps

B. Experimental Results

The feature vectors of all the reference speech samples are stored in the database in the training phase. In the matching phase, the test sample that is to be identified is taken and similarly processed as in the training phase to form the feature vector. The stored feature vector which gives the minimum Euclidean distance with the input sample feature vector is declared as the speaker identified.

Table II gives the number of matches for the four different transforms. The matching has been calculated by considering the minimum Euclidean distance between the feature vector of the test speech signal and the feature vector of the speech signals stored in the database. The rows of Table II show the number of samples of each speech signal grouped together to form the columns of a matrix whose transform is then taken. For each grouping, the transform which gives maximum matches has been shaded in yellow. We can see that for groupings of 64, 128 and 256 DST gives the best matching i.e. 86, 98 and 99 (out of 105) respectively. For a grouping of 512, DCT gives best matching i.e. 99. For a grouping of 1024 samples, DST gives maximum matches i.e. 101. It can also be seen that as the number of samples grouped is further increased beyond 1024, the number of matches is reduced for all the transforms.

TABLE II. NO. OF MATCHES FOR DIFFERENT GROUPINGS

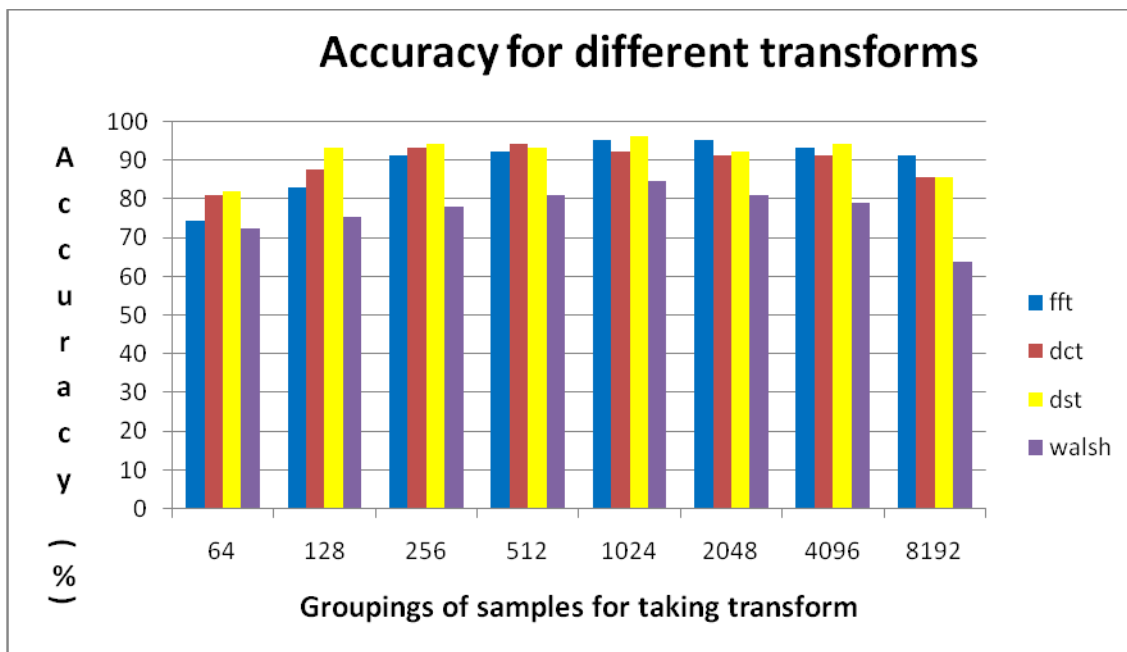
No. of samples grouped	Number of matches (out of 105)			
	FFT	DCT	DST	WALSH
64	78	85	86	76
128	87	92	98	79
256	96	98	99	82
512	97	99	98	85
1024	100	97	101	89
2048	100	96	97	85
4096	98	96	99	83
8192	96	90	90	67

C. Accuracy of Identification

The accuracy of the identification system is calculated as given by equation 5.

$$\text{Accuracy (\%)} = \frac{\text{number of matches}}{\text{number of samples tested}} \quad (5)$$

The accuracy for the different groupings of the four transforms was calculated and is shown in Figure 3.



The results show the accuracy increases as we increase the feature vector size from 64 to 512 for the transforms. Only for DST, the accuracy decreases from 94.28% to 93.33% as we increase the feature vector size from 256 to 512. The feature vector size of 1024 gives the best result for all the transforms except DCT. For DCT, the best result is obtained for a feature vector size of 512. For DFT, the maximum accuracy obtained is 95.2381% for a feature vector size of 1024. Walsh transform gives a maximum accuracy of around 84.7619%. DST performs best giving a maximum accuracy of 96.1905% for a feature vector size of 1024.

V. CONCLUSION

In this paper we have compared the performance of four different transforms for speaker identification. All the Transforms give an accuracy of more than 80% for the feature vector size considered. Accuracy increases as the feature vector size is increased from 64 onwards. But for feature vector size of more than 1024 the accuracy again starts decreasing. The results show that DST performs best. The maximum accuracy obtained for DST is around 96% for a feature vector size of 1024. The present study is ongoing and we are analyzing the performance on other transforms.

REFERENCES

- [1] Lawrence Rabiner, Biing-Hwang Juang and B.Yegnanarayana, "Fundamental of Speech Recognition", Prentice-Hall, Englewood Cliffs, 2009.
- [2] S Furui, "50 years of progress in speech and speaker recognition research", ECTI Transactions on Computer and Information Technology, Vol. 1, No. 2, November 2005.
- [3] D. A. Reynolds, "An overview of automatic speaker recognition technology," Proc. IEEE Int. Conf. Acoust., Speech, S
- [4] Joseph P. Campbell, Jr., Senior Member, IEEE, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.
- [5] S. Furui. Recent advances in speaker recognition. AVBPA97, pp 237--251, 1997
- [6] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovsk-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," EURASIP J. Appl. Signal Process., vol. 2004, no. 1, pp. 430-451, 2004.
- [7] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639-643, Oct. 1994.
- [8] Tomi Kinnunen, Evgeny Karpov, and Pasi Fränti, "Realtime Speaker Identification", ICSP2004.
- [9] Marco Grimaldi and Fred Cummins, "Speaker Identification using Instantaneous Frequencies", IEEE Transactions on Audio, Speech, and Language Processing, vol., 16, no. 6, August 2008.
- [10] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). "Binary Quantization of Feature vectors for robust text-independent Speaker Identification" in IEEE Transactions on Speech and Audio Processing Vol. 7, No. 1, January 1999. IEEE, New York, NY, U.S.A.
- [11] H B Kekre, Vaishali Kulkarni, "Speaker Identification by using Vector Quantization", International Journal of Engineering Science and Technology, May 2010.

- [12] H B Kekre, Vaishali Kulkarni, "Performance Comparison of Speaker Recognition using Vector Quantization by LBG and KFCG" , International Journal of Computer Applications, vol. 3, July 2010.
- [13] H B Kekre, Vaishali Kulkarni, "Performance Comparison of Automatic Speaker Recognition using Vector Quantization by LBG KFCG and KMCG" , International Journal of Computer Science and Security, Vol: 4 Issue: 4, 2010.
- [14] H B Kekre, Vaishali Kulkarni, "Comparative Analysis of Automatic Speaker Recognition using Kekre's Fast Codebook Generation Algorithm in Time Domain and Transform Domain" , International Journal of Computer Applications, Volume 7 No.1. September 2010.
- [15] Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo "Performance Comparison of Image Retrieval using Row Mean of Transformed Column Image", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1908-1912
- [16] Dr.H.B.Kekre,Sudeep Thepade "Edge Texture Based CBIR using Row Mean of Transformed Column Gradient Image", International Journal of Computer Applications (0975 - 8887) Volume 7- No.10, October 2010
- [17] Dr. H.B.Kekre, Sudeep D. Thepade, Akshay Maloo "Eigenvectors of Covariance Matrix using Row Mean and Column Mean Sequences for Face Recognition", International Journal of Biometrics and Bioinformatics (IJBB), Volume (4): Issue (2)
- [18] Dr. H.B.Kekre, Sudeep Thepade, Archana Athawale, "Grayscale Image Retrieval using DCT on Row mean, Column mean and Combination", Journal of Sci., Engg. & Tech. Mgt. Vol 2 (1), January 2010
- [19] Dr. H. B. Kekre, Dr. T. K. Sarode, Shachi J. Natu, Prachi J. Natu "Performance Comparison of Speaker Identification Using DCT, Walsh, Haar on Full and Row Mean of Spectrogram", International Journal of Computer Applications (0975 - 8887) Volume 5- No.6, August 2010
- [20] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform", IEEE Trans. Computers, 90-93, Jan 1974.
- [21] N. Ahmed, "How I came up with the Discrete Cosine Transform", Digital Signal Processing, Vol. 1, 1991.
- [22] G. Strang, "The Discrete Cosine Transform," SIAM Review, Volume 41, Number 1, 1999.
- [23] Fino, B.J., and Algazi, V.R., 1976, "Unified Matrix Treatment of the Fast Walsh-Hadamard Transform," IEEE Transactions on Computers 25: 1142-1146

AUTHORS PROFILE



Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engg. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970. He has worked Over 35 years as Faculty of Electrical Engineering and then HOD Computer Science and Engg. at IIT Bombay. For last 13 years worked as a Professor in Department of Computer Engg. at Thadomal Shahani Engineering College, Mumbai. He is currently Senior Professor working with Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS University, Vile Parle(w), Mumbai, INDIA. He has guided 17 Ph.D.s, 150 M.E./M.Tech Projects and several B.E./B.Tech Projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networks. He has more than 300 papers in National / International Conferences / Journals to his credit. Recently twelve students working under his guidance have received best paper awards. Recently two research scholars have received Ph. D. degree from NMIMS University Currently he is guiding ten Ph.D. students. He is member of ISTE and IETE.

Vaishali Kulkarni has received B.E in Electronics Engg. from Mumbai University in 1997, M.E (Electronics and Telecom) from Mumbai University in 2006. Presently she is pursuing Ph. D from NMIMS University. She has a teaching experience of more than 8 years. She is Associate Professor in telecom Department in MPSTME, NMIMS University. Her areas of interest include Speech processing: Speech and Speaker Recognition. She has 8 papers in National / International Conferences / Journals to her credit.



Performance Evaluation of Space-Time Turbo Code Concatenated With Block Code MC-CDMA Systems

Lokesh Kumar Bansal

Department of Electronics & Comm. Engg.,
N.I.E.M., Mathura, India
e-mail: lokesh_bansal@rediffmail.com

Aditya Trivedi

Department of Information and Comm. Technology
ABV-IIITM, Gwalior, India
e-mail: atrivedi@iiitm.ac.in

Abstract—In this paper, performance of a space-time turbo code (STTuC) in concatenation with space-time block code (STBC) in multi-carrier code-division multiple-access (MC-CDMA) system with multi-path fading channel is considered. The performance in terms of bit error rate (BER) is evaluated through simulations. The corresponding BER of the concatenated STTuC-STBC-MC-CDMA system is compared with STTuC-MC-CDMA system and STBC-MC-CDMA system. The simulation results show that the STTuC-MC-CDMA system performance is better the STBC-MC-CDMA system and it can be further improved by employing concatenation technique i.e. STTuC-STBC-MC-CDMA system.

Keywords—MMSE; Multi-path channel; MAP Decoder; MIMO; Space-time code; Space-time turbo-code; Space-time trellis-code; Viterbi Decoder

I. INTRODUCTION

The wireless communication market is increased exponentially in recent years. A lot of interest has been developed in modulation techniques like Orthogonal Frequency Division Multiplexing (OFDM), Code Division Multiple Access (CDMA) and Multicarrier Code Division Multiple Access (MC-CDMA). MC-CDMA is seen as a possible candidate for Fourth Generation (4G) wireless communication systems that demand higher data rate for voice and data transmissions. CDMA technique is widely used in current Third Generation (3G) wireless communication systems. The principle of spread spectrum technology behind CDMA was popularly used in military communications for improving secrecy and low probability of interception during transmission. Now, CDMA technology is also increased in civilian markets due to high capacity and better performance. The rapid growth of video, voice and data transmission through the internet and the increased use of mobile telephony in today's life have the necessity for higher data rate transmissions over the wireless channels [1]-[4].

In 3G systems we have higher data rate i.e. 64kbps – 2Mbps as compared to 9.6kbps – 14.4kbps used in 2G systems. The 4G systems that include broadband wireless services require data rate up to 20Mbps. This also

emphasizes the need for improved spectral efficiency and higher Quality of Service (QoS) over current systems [5]-[7]. The above requirements can be fulfilled by multicarrier modulation techniques. Single carrier systems give good data rate but are limited in performance in multi path fading channels. Improved performance in multipath fading channel conditions, high data rates and efficient bandwidth usage are the main advantages of multicarrier modulation. Space-time coding (STC) techniques incorporate the methods of transmitter diversity, channel coding, and provide significant capacity gains over the traditional communication systems in the fading wireless channels. Here, STC has been developed along two major directions: *space-time block coding* (STBC) and *space-time turbo coding* (STTuC).

In this paper, the STBC, STTuC, and STTuC-STBC code techniques are studied and applied in MC-CDMA systems. These techniques are employed with multiple input multiple output (MIMO) antenna diversity in multi-path fading channel. At the receiver side minimum mean-square error detection (MMSE) technique is used by employing maximum a posteriori (MAP) algorithm for turbo code decoding purpose on full load. The performance in terms of bit error rate probability (BER) is obtained in presence of perfect channel state information (CSI).

The rest of the paper is organized as follows: in Section II, MC-CDMA system is presented. In Section III, space-time coding technique is described. The space-time block code scheme is given in section IV. In Section V, the mathematical representation for the space-time turbo code is explained. Space-Time turbo code in concatenation with space-time block code MC-CDMA system model is given in section VI. Simulation for error rate performance of MC-CDMA systems in presence of perfect CSI are carried out in Section VII. The conclusions are presented in Section VIII.

II. MC-CDMA SYSTEMS

CDMA communication system allows multiple users to transmit at the same time in the same frequency band. Traditional multiple access techniques like time division multiple access (TDMA) and frequency division multiple access (FDMA) are based on the philosophy of letting no more than one transmitter occupy a given time-frequency

slot. In a CDMA based system, users are assigned different signature wave forms or codes. Each transmitter sends its data stream by modulating its own signature waveform as in a single-user digital communication system. The receiver does not need to concern itself with the fact that the signature waveforms overlap both in frequency and time, because their orthogonality ensures that they will be transparent to the output of the other user's correlator [8], [9]. CDMA still has a few drawbacks, the main one being that capacity is limited by the multiple access interference (MAI). The W-CDMA supports high data rate transmission, typically 384 kbps for wide area coverage and 2 mbps for local coverage for multimedia services. Thus, W-CDMA is capable of offering the transmission of voice, text, data, picture (still image) and video over a single platform. However, in addition to the drawbacks arising from the mobile environment and multiple access interference, high bit rate transmission causes inter-symbol interference (ISI) to occur. The ISI, therefore, has to be taken into account during transmission.

Multi-carrier modulation is being proposed for 4G wireless communication systems for high data rate application to reduce the effect of ISI and adapt to channel conditions. A number of MC-CDMA systems have been proposed lately. These systems solve the ISI problem by transmitting the same data symbol over a large number of narrow band orthogonal carriers. The number of carriers equals or exceeds the pseudo-noise (PN) code length [10]-[12]. In MC-CDMA system, each data symbol is transmitted over N narrowband sub-carriers, where each sub-carrier is encoded with a 0 or π phase offset. An MC-CDMA signal is composed of N narrowband sub-carrier signals each with symbol duration, T_b , much larger than the delay spread, T_d , hence MC-CDMA signal does not experience significant ISI. Multiple access is achieved with different users transmitting at the same set of sub-carriers but with spreading codes that are different to the codes of other users. Initially, the data stream is serial to parallel converted to a number of lower rate streams. Each stream feeds a number of parallel streams with the same rate. On each of the parallel streams, bits are interleaved and spread by a PN code with a suitable chip rate. Then, these streams modulate different or orthogonal carriers with a successively overlapping bandwidth.

In a MC-CDMA system, the numbers of carriers are typically chosen to be large enough so that the signal on each sub-carrier is propagated through a channel, which behaves in a nonselective manner. The fading processes on each sub-carrier for each user must be estimated, which can be used in forming the MMSE filter. This approach is shown to perform close to ideal for sufficiently high vehicle speeds, up to which the normalized Doppler rate is about 1percent [13].

Providing high data rate transmission of the order of several megabits per second (mbps) is important for future wireless communications. In recent years, antenna systems which employ multiple antennas at both the base station (BS) and mobile station (MS), have been proposed and

demonstrated to significantly increase system performance as well as capacity. The merit of using multiple antennas or space diversity is that no bandwidth expansion or increase in transmitted power is required for capacity and performance improvements [14].

III. S-T CODING TECHNIQUE

In most wireless communication systems, the number of diversity methods are used to get the required performance. According to the domain, the diversity techniques are classified into time, frequency, and space diversity [15]. S-T coding technique is designed for use with multiple transmit antennas. There are various techniques in coding structures, which include Alamouti STC, STBC, STTC, STTuC, and layered space-time (LST) codes. S-T coding with multiple transmit and receive antennas minimizes the effect of multipath fading and improves the performance and capacity of digital transmission over wireless radio channels [16].

STBC can achieve a maximum possible diversity advantage with a simple decoding algorithm. It is very attractive because of its simplicity. However, no coding gain can be provided by STBC. STTuC is able to combat the effects of fading. However, STTuC have a potential drawback due to the fact that its decoder complexity (MAP decoder) grows with the number of iterations.

A base band S-T coded system with n_T transmit antennas and n_R receive antennas is shown in Figure 1. The transmitted data are encoded by a S-T encoder. At each time instant, a block of m binary information symbols, denoted by $\mathbf{a}(n) = (a^1(n), a^2(n), a^3(n), \dots, a^m(n))$ is fed into the S-T encoder. The S-T encoder maps the block of m binary input data into n_T modulation symbols from a signal set of $M = 2^m$ points. The coded data are applied to a serial to parallel (S/P) converter producing a sequence of n_T parallel symbols, arranged into a $n_T \times 1$ column vector

$$\mathbf{X}(n) = (x^1(n), x^2(n), \dots, x^{n_T}(n))^T$$

where T denotes the transpose of a matrix. The n_T parallel outputs are simultaneously transmitted by n_T different antennas, where by symbol $x^i(n)$, $1 \leq i \leq n_T$, is transmitted by antenna i and all transmitted symbols have the same duration of T sec. The vector of coded modulation symbols from different antennas is called a S-T symbol. The

spectral efficiency of the system is $\eta = \frac{r_b}{B} = m$

bits/sec/Hz, where r_b is the data rate and B is the channel bandwidth. This spectral efficiency is equal to the spectral efficiency of a reference uncoded system with one transmit antenna.

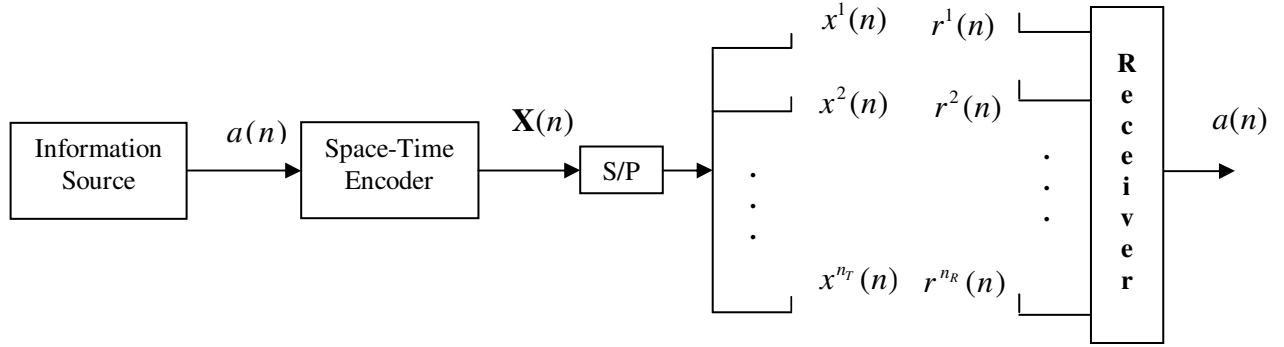


Figure 1- A base band system model

The multiple antennas at both the transmitter and the receiver create a MIMO channel. For wireless mobile communications, each link from a transmit antenna to a receive antenna can be modeled by flat fading, if we assume that the channel is memoryless. The MIMO channel with n_T transmit and n_R receive antennas can be represented by an $(n_R \times n_T)$ channel matrix $\mathbf{H}$ and it can be given as

$$\mathbf{H}(n) = \begin{bmatrix} h_{1,1}(n) & h_{1,2}(n) & \dots & h_{1,n_T}(n) \\ h_{2,1}(n) & h_{2,2}(n) & \dots & h_{2,n_T}(n) \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_R,1}(n) & h_{n_R,2}(n) & \dots & h_{n_R,n_T}(n) \end{bmatrix}$$

where the $(ji)^{th}$ element, denoted by $h_{j,i}(n)$, is the fading attenuation coefficient for the path from transmit antenna i to receive antenna j .

It is further assumed that the fading coefficients $h_{j,i}(n)$ are independent complex Gaussian random variables. At the receiver, the signal at each of the n_R receive antennas is a noisy superposition of the n_T transmitted signals degraded by channel fading. The n^{th} received signal at antenna j ($j = 1, 2, \dots, n_R$) denoted by $r^j(n)$, is given by

$$r^j(n) = \sum_{i=1}^{n_T} h_{j,i}(n)x^i(n) + v^j(n) \quad (1)$$

where $v^j(n)$ is the noise component of receive antenna j at time n , which is an independent noise sample of the zero-mean complex Gaussian random variable with the one sided power spectral density of N_o . $r(n)$ is the received signal sequence from n_R receive antennas of $n_R \times 1$ column vector

$$\mathbf{r}(n) = (r^1(n), r^2(n), \dots, r^{n_R}(n))^T$$

Thus, the received signal vector can be represented as,

$$\mathbf{r}(n) = \mathbf{H}(n)\mathbf{X}(n) + \mathbf{v}(n) \quad (2)$$

It is assumed that the decoder at the receiver uses a maximum likelihood algorithm to estimate the transmitted information sequence with receiver has perfect channel state information (CSI) on the MIMO channel. At the receiver, the decision metric is computed based on the squared Euclidean distance between the hypothesized received sequence and the actual received sequence as

$$\sum_{j=1}^{n_R} \left| r^j(n) - \sum_{i=1}^{n_T} h_{j,i}(n)x^i(n) \right|^2 \quad (3)$$

The decoder select a codeword with the minimum decision metric as the decoded sequence [4], [16].

IV. SPACE-TIME BLOCK CODE (STBC)

STBC first introduced by Alamouti with two transmit antennas. Figure 2 shows the block diagram of the Alamouti S-T encoder. It is assumed that a M -ary modulation scheme is used. In the Alamouti S-T encoder, each group of m information bits is first coded, where $m = \log_2 M$. Here, the encoder takes a block of two modulated symbols x_1 and x_2 in each encoding operation and maps them to the transmit antennas according to a code matrix given by

$$\mathbf{X} = \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} \quad (4)$$

The encoder outputs are transmitted in two consecutive transmission periods from two transmit antennas. During the first transmission period, two signals x_1 and x_2 are transmitted simultaneously from antenna one and antenna two, respectively.

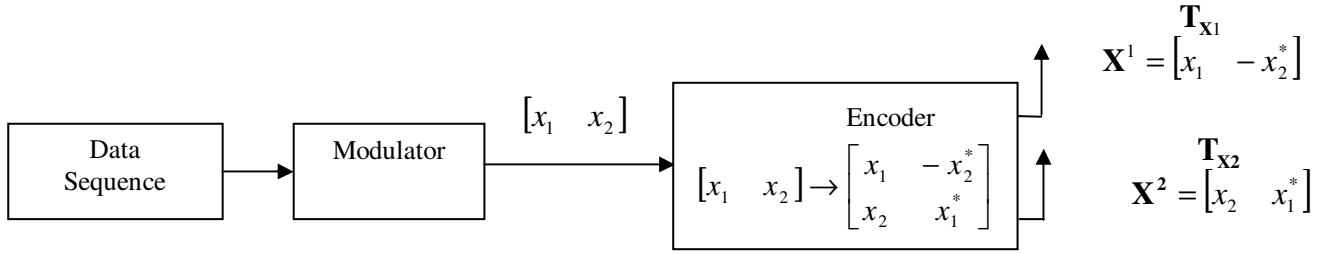


Figure 2- Block diagram of the Alamouti space-time encoder

In the second transmission period, signal $-x_2^*$ is transmitted from transmit antenna one and signal x_1^* from transmit antenna two [7], [16]. The main principle of the Alamouti scheme is that the transmit sequences from the two transmit antennas are orthogonal, since the inner product of the sequences $\mathbf{x}^1$ and $\mathbf{x}^2$ is zero, i.e.

$$\mathbf{x}^1 \cdot \mathbf{x}^2 = x_1 \cdot x_2^* - x_2^* \cdot x_1 = 0$$

This scheme may be generalized to an arbitrary number of transmit antennas by applying the theory of orthogonal designs. The generalized schemes are referred to as STBC. The STBC can achieve the full transmit diversity specified by the number of the transmit antennas n_T , while allowing a very simple maximum-likelihood decoding algorithm, based only on linear processing of the received signals [8], [16].

In general, a STBC is defined by a $n_T \times p$ transmission matrix $\mathbf{X}$. Here n_T represents the number of transmit antennas and p represents the number of time periods for transmission of one block coded symbols. It is assumed that the signal constellation consists of 2^m points. At each encoding operation, a block of km information bits are mapped into the signal constellation to select k modulated signals $x_1, x_2, \dots, x_k$, where each group of m bits selects a constellation signal. The k modulated signals are encoded by a S-T block encoder to generate n_T parallel signal sequences of length p according to the transmission matrix $\mathbf{X}$. These sequences are transmitted through n_T transmit antennas simultaneously in p time periods.

$$\mathbf{X} \cdot \mathbf{X}^H = \alpha \left(|x_1|^2 + |x_2|^2 + \dots + |x_k|^2 \right) \mathbf{I}_{n_T}$$

where α is a constant, $\mathbf{X}^H$ is the Hermitian of $\mathbf{X}$ and $\mathbf{I}_{n_T}$ is an $n_T \times n_T$ identity matrix. The element of $\mathbf{X}$ in the i^{th} row and j^{th} column, $x_{i,j}$, $i = 1, 2, \dots, n_T$, $j = 1, 2, \dots, p$, represents the signal transmitted from the antenna i at time j . The orthogonal designs are applied to construct

STBC. The rows of the transmission matrix $\mathbf{X}_{n_T}$ are orthogonal to each other. This means that in each block, the signal sequences from any two transmit antennas are orthogonal. For example, if we assume that $\mathbf{X}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ is the transmitted sequence from the i^{th} antenna, $i = 1, 2, \dots, n_T$, we have

$$\mathbf{X}_i \cdot \mathbf{X}_j = \sum_{t=1}^p x_{i,t} \cdot x_{j,t}^* = 0,$$

$$i \neq j, i, j \in \{1, 2, \dots, n_T\}$$

where $\mathbf{X}_i \cdot \mathbf{X}_j$ denotes the inner product of the sequences $\mathbf{X}_i$ and $\mathbf{X}_j$. The orthogonality enables to achieve the full transmit diversity for a given number of transmit antennas. In addition, it allows the receiver to decouple the signals transmitted from different antennas and consequently, a simple maximum likelihood decoding, based only on linear processing of the received signals [8], [16].

V. SPACE-TIME TURBO CODE (STTC)

Space-Time Turbo code can achieve outstanding performance gain because it uses specific coding and decoding structure. Here, in turbo code parallel concatenation convolutional code is used, which is a recent scheme of channel code developed, whose bit error performance is close to the limit predicted by Shannon. In coding side it uses a particular concatenation of two Recursive Convolutional Codes (RCC), associated together by the inter-leaver, while in decoding it delivered soft information between two decoder components to make soft-input and soft-output iterative decoding. The first decoder must deliver a weighted soft decision to the second decoder. The Logarithm of Likelihood Ratio (LLR) $L_j(a_n)$ associated with each decoded bit a_n by first decoder is a relevant piece of information for the second decoder. For information bit a_n , It's LLR is defined as the natural log of its base probabilities.

$$L(a_n) = \ln \frac{P(a_n = +1)}{P(a_n = -1)}$$

If a_n has two values +1 and -1, with equally likely, then this ratio is equal to zero. If the binary variable is not equally likely then $P(a_n = -1) = 1 - P(a_n = +1)$.

$$L(a_n) = \ln \frac{P(a_n = +1)}{1 - P(a_n = +1)}$$

The LLR of N bit sequence is formulated as below. The lower indicator of y means it starts at time 1 and the upper indicator means the ending point at N .

$$L_1(a_n) = \ln \frac{P(a_n = +1 | y_1^N)}{P(a_n = -1 | y_1^N)} \quad (5)$$

Here, first the S-T Turbo encoder encodes the source data. Next, the encoded data is applied to interleaver, and then mapped according to the desired signal constellation. Finally at each time interval, the signals are modulated and transmitted simultaneously over different transmit antennas [16]-[18]. Using Baye's rule, the above equation can be reformulated as

$$L(a_n) = \ln \frac{P(y_1^N, a_n = +1) / P(y_1^N)}{P(y_1^N, a_n = -1) / P(y_1^N)} = \ln \frac{P(y_1^N, a_n = +1)}{P(y_1^N, a_n = -1)}$$

This LLR includes joint probabilities between the received bits and the information bit, the numerator of which can be written as

$$\sum_N P(a_n = +1, y_1^N) = \sum_N P(s', s, y_1^N)$$

where s' is starting state and s is ending state of trellis.

$$L(a_n) = \ln \frac{P(y_1^N, a_n = +1)}{P(y_1^N, a_n = -1)} = \frac{\sum_{a_n=+1} P(s', s, y_1^N)}{\sum_{a_n=-1} P(s', s, y_1^N)}$$

$$y_1^N = y_1^{k-1}, y_n, y_{n+1}^N = y_p, y_k, y_f$$

We take the N bit data sequence and separate this into three pieces, from 1 to $n-1$, then the n^{th} point, and then from $n+1$ to N . $P(s', s, y_1^N) = P(s', s, y_p, y_k, y_f)$ where y_p is past sequence, which is the part that came before the current data point. y_n current data point and y_f is the part that come after the current point. Using Baye's rule

$$\begin{aligned} P(s', s, y) &= P(s', s, y_p, y_n, y_f) \\ &= P(y_f | s', s, y_p, y_n) P(s', s, y_p, y_n) \end{aligned}$$

The term y_f is the future sequence and we consider that it is independent of the past and only depend on the present state s . We can remove these dependencies to simplify the above equation.

$$P(s', s, y) = P(y_f | s) P(s', s, y_p, y_n).$$

Now apply Baye's rule to the last term

$$P(s', s, y_p, y_n) = P(s, y_n | s', y_p) P(s', y_p)$$

$$P(s', s, y) = P(y_f | s) P(s, y_n | s', y_p) P(s', y_p)$$

$$\text{Let } \alpha_{n-1}(s') = P(s', y_p), \beta_n(s) = P(y_f | s),$$

$$\text{and } \gamma_n(s', s) = P(s, y_n | s', y_p)$$

$$\text{Then } P(s', s, y) = \alpha_{n-1}(s') \beta_n(s) \gamma_n(s', s)$$

The LLR equation for MAP algorithm can be written as

$$L(a_n) = \frac{\sum_{a_n=+1} \alpha_{n-1}(s') \beta_n(s) \gamma_n(s', s)}{\sum_{a_n=-1} \alpha_{n-1}(s') \beta_n(s) \gamma_n(s', s)}, \text{ where } \alpha_{n-1}(s') \text{ is called}$$

the Forward metric, $\beta_n(s)$ is called the Backward metric,

and $\gamma_n(s', s)$ is called Transition metric. At the receiver, the received data sequence is combined according to the combining techniques described for STC-MC-CDMA system. The soft output of the combiner is applied directly to the deinterleaver, and then finally, it is applied to S-T Turbo decoder, such as the MAP algorithm, to decode the data.

$$L_1(a_n) = L(a - \text{priori}) + L(\text{channel}) + L(\text{extrinsic})$$

The $L(a - \text{priori})$ is a-priori information about a_n ,

$L(\text{channel})$ is the channel value calculated from the knowledge of the SNR and received signal. The third term is called the a-posteriori term, also called the extrinsic L value.

$$L(\text{extrinsic}) = \ln \frac{\sum_{a'} \bar{\alpha}_{n-1}(s') \bar{\beta}_n(s) \gamma_n^e(s', s)}{\sum_{a'} \bar{\alpha}_{n-1}(s') \bar{\beta}_n(s) \gamma_n^e(s', s)} \quad (6)$$

During each iteration the decoder produces the extrinsic value, this extrinsic value becomes the input to the next decoder. The decision is made about the bit by looking at the sign of the L value. $\bar{a}_n = \text{sign}\{L_1(a_n)\}$, The process can continue until the extrinsic values change becomes insignificant or the algorithm can allow for a fixed number of iterations.

VI. STTuC CONCATENATED WITH STBC MC-CDMA SYSTEM MODEL

In order to further improvement in the performance of STTuC-MC-CDMA system, we can use S-T turbo code in concatenation with S-T block code MC-CDMA system. The STTuC-STBC-MC-CDMA system provides both diversity and coding gain with a reasonable increase in complexity. Figure 3 show the general block diagram of concatenated system. First, the STTuC encoder encodes the source data. Next, the encoded data is applied to S-T block encoder &

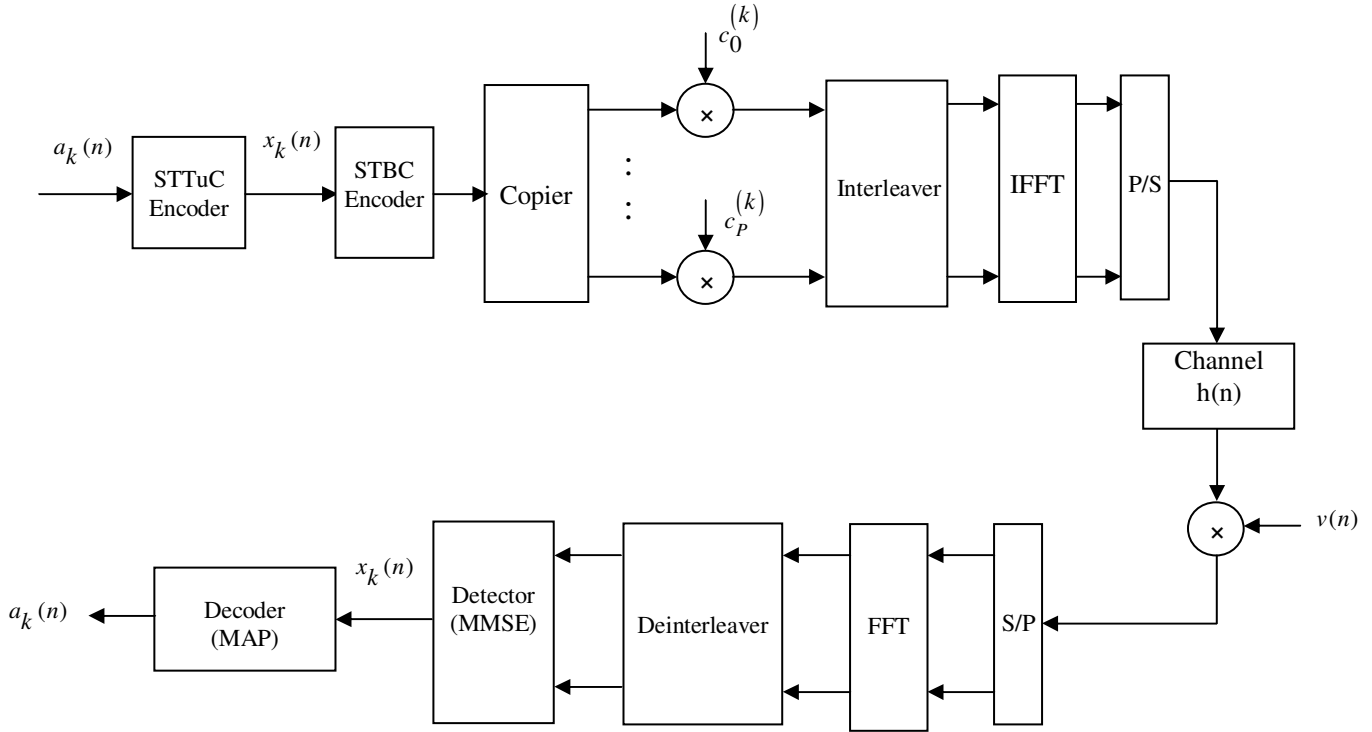


Figure 3- Block diagram of STTuC Concatenated with STBC code MC-CDMA System

interleaver, and then mapped according to the desired signal constellation. Finally at each time interval, the symbols are modulated and transmitted simultaneously over different transmit antennas. At the receiver, the received data is combined according to the combining techniques described for STBC. The soft output of the combiner is sent directly to the deinterleaver, and then finally, it is applied to a STTuC decoder, such as the MAP algorithm, to decode the data.

Here, a base band system configuration of the STBC-MC-CMDA system employing the Alamouti's S-T coding scheme at the transmitter is depicted in Figure 3, which involves two transmit antennas, Tx1 and Tx2, and one receive antenna, Rx. At the transmitter, we assume K number of users transmit simultaneously with STTuC in concatenation of STBC over MC-CDMA system from the two transmit antennas. The frequency selective channel between transmit and receive antennas is divided into P subchannels such that each subchannel is approximately flat. Let $\{X^{(k)}(n)\}$ be the S-T Turbo encoded output.

For k^{th} user, the output of S-T Turbo encoder is given to the S-T block encoder that is represented by the following code matrix

$$S^{(k)}(n) = \begin{bmatrix} s^{(k)}(1) & s^{(k)}(2) \\ \bar{s}^{(k)}(1) & \bar{s}^{(k)}(2) \end{bmatrix} \quad (7)$$

where

$$\begin{aligned} s^{(k)}(1) &= x^{(k)}(1), s^{(k)}(2) = -x^{(k)*}(2) \\ \bar{s}^{(k)}(1) &= x^{(k)}(2), \bar{s}^{(k)}(2) = x^{(k)*}(1) \end{aligned}$$

and $(.)^*$ denotes the complex conjugate. The two columns of $s^{(k)}(n)$ will be transmitted in two consecutive time slots, with the first element of each column transmitted from Tx1 and the second element from Tx2, respectively. Throughout this paper, $(\bar{\cdot})$ is designated to quantities associated with Tx2. In the current system, each user is assigned two distinct spreading codes to spread symbols transmitted from the two antennas. Let $c^{(k)} = [c_0^{(k)}, \dots, c_{p-1}^{(k)}]^T$ and

$\bar{c}^{(k)} = [\bar{c}_0^{(k)}, \dots, \bar{c}_{p-1}^{(k)}]^T$ are two spreading code sequences for k^{th} user with processing gain P which spread the symbols transmitted from antenna Tx1 & Tx2 respectively. Define $u^{(k)}(n) = s^{(k)}(n)c^{(k)}$ is the signal associated with Tx1. The OFDM modulation can be implemented via IFFT on $u^{(k)}(n)$

$$y^{(k)}(n) = F^{-1}u^{(k)}(n) \quad (8)$$

Similarly we obtain the signal associated with Tx2 as $\bar{y}^{(k)}(n)$.

At the receiver, the received signal after demodulating via FFT and passing through deinterleaver is given as

$$Y(n) = \sum_{k=1}^K \left[s^{(k)}(n)bc^{(k)} + \bar{s}^{(k)}(n)\bar{b}\bar{c}^{(k)} \right] + V(n)$$

where $b = Fh$, and $h = [h(0), \dots, h(M-1), 0_{P-M}^T]^T$ is channel vector between Tx1 & Rx, and $v(n) = [v_0(n), \dots, v_{p-1}(n)]^T$ contains samples of the channel noise with zero mean and variance σ_v^2 .

VII. SIMULATION RESULTS

The simulations are done for STTuC-MC-CDMA, STBC-MC-CDMA, and STTuC-STBC-MC-CDMA systems with $K = 10$ users. The user symbols are drawn from a unit-energy BPSK (binary phase shift keying) constellation. Walsh-Hadamard codes with processing gain $P = 32$ are used for spreading. We assume noise samples as i.i.d. complex Gaussian random variables with zero mean and variance σ_n^2 .

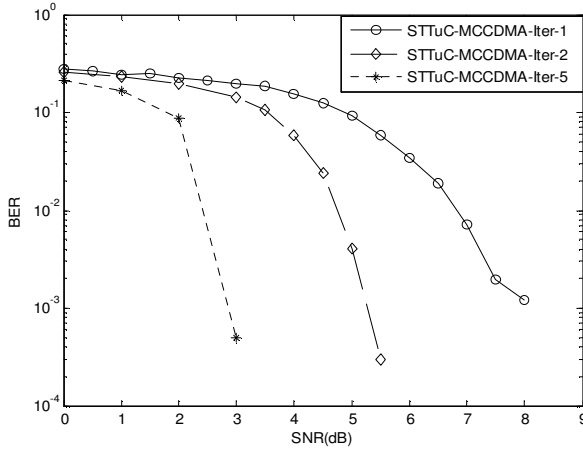


Figure 4- BER performance of STTuC-MC-CDMA System for 1, 2, & 5 number of iterations with perfect CSI.

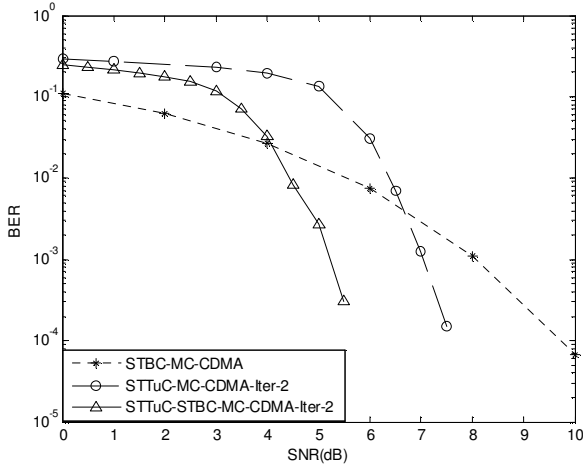


Figure 5- Performance comparison of STTuC-STBC, STTuC, and STBC-MC-CDMA Systems with perfect CSI.

In Figure 4 we present the BER of STTuC-MC-CDMA system with different number of iterations, and it is observed that the performance is improved by 2 dB from 1 to 2 iterations and again it is enhanced by around 2 dB from 2 to 5 iterations at 10^{-3} BER. From this we can state that the rate of improvement in the performance decreases with increasing number of iterations. The performance of STBC-MC-CDMA, STTuC-MC-CDMA, & STTuC-STBC-MC-CDMA systems versus the SNR in a Rayleigh fading environment for $K = 10$ users and 20000 bit sequences are shown in Figure 5. It is noted that at 10^{-3} BER the performance of STTuC-MC-CDMA system using MMSE detector is better than STBC-MC-CDMA system by around 1 dB. The performance of STTuC-MC-CDMA system can also be further improved by 1.5dB at 10^{-3} BER by using STTuC-STBC-MC-CDMA systems in presence of perfect channel state information (CSI).

VIII. CONCLUSION

In this paper, the performance of concatenated STTuC-STBC-MC-CDMA, STTuC-MC-CDMA and STBC-MC-CDMA systems are obtained using MMSE detection technique employing MAP algorithm for turbo code decoding. Simulation results are presented in presence of perfect CSI. It is noted that the performance of STTuC-MC-CDMA system increases with more number of iterations. However, the rate of improvement in the performance decreases with increasing number of iterations. It is observed that the performance of STTuC-MC-CDMA system with 2 numbers of iterations is better than STBC-MC-CDMA system by around 1 dB. The performance gain can be further improved by around 1.5 dB using STTuC-STBC-MC-CDMA system at 10^{-3} BER.

REFERENCES

- [1] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," IEEE Transactions on Information Theory, vol.44, no. 2, pp. 744-765, March 1998.
- [2] Hemanth Sampath, and S. Talwar, J. Tellado, "A Fourth-Generation MIMO-OFDM Broadband Wireless System: Design, Performance, and Field Trial Results," IEEE Comm. Magazine, pp.143-149, September 2002.
- [3] M. Juntti, M. Vehkaperä, J. Leinonen, "MIMO MC-CDMA Communications for Future Cellular Systems," IEEE Comm. Magazine, pp. 118-124, Feb.2005.
- [4] S. Hijazi, B. Natarajan, and Z. Wu., "Flexible Spectrum use and Better Coexistence at the Physical Layer of Future Wireless Systems via a Multicarrier Platform", IEEE Wireless Comm., April 2004, pp. 64-71.
- [5] Essam A. Sourour, and Masao Nakagawa, "Performance of Orthogonal Multicarrier CDMA in a Multipath Fading Channel," IEEE Transactions on Communications, vol. 44, no.3, pp. 356-366, March 1996.
- [6] S. Hara, and R. Prasad, "An Overview of Multi-Carrier CDMA," IEEE Comm. Magazine, vol. 35, no. 12, pp.126-133, Dec. 1997.
- [7] Scott L. Miller, and B. J. Rainbolt, "MMSE Detection of Multicarrier CDMA," IEEE J. on Selected Areas in Communications, vol. 18, no.11, pp. 2356-2362, Nov. 2000.
- [8] Kai-Kit Wong, Ross D. Murch, and Khaled Ben Letaief, "Performance Enhancement of Multiuser MIMO Wireless Communication Systems," IEEE Transactions on Communications, vol. 50, no.12, pp. 1960-1968, Dec. 2002.
- [9] Proakis, J. G. (2008). Digital Communication (5th ed.). New York: McGraw Hill.

- [10] A. F. Naguib, V. Tarokh, N. Seshadri, and A. R. Calderbank, "A space-time coding modem for high-data-rate wireless communications," *IEEE J. on Selected Areas in Comm.*, vol. 16, no. 8, pp. 1459 – 1478, Oct. 1998.
- [11] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451 – 1458, Oct. 1998.
- [12] Trivedi Aditya and Bansal Lokesh Kumar, "Comparative study of different space-time coding schemes for MC-CDMA systems", *International Journal of Communication, Network and System Sciences* published by Scientific Research, vol. 3, no. 3, pp. 418-424, April 2010.
- [13] Wei Sun, Hongbin Li, and Moeness Amin, "A Subspace-Based Channel Identification Algorithm for Forward Link in Space-Time Coded MC-CDMA Systems", *IEEE Proceedings*, 2002 pp. 445-448.
- [14] L. A. P. Hernandez and M. G. Otero, "A new STB-TCM coded MC-CDMA Systems with MMSE-SVA based Decoding and Soft-Interference Cancellation", *IEEE Proceedings* 2005, pp. 113-116.
- [15] Clude Berrou, Alain Glavieux, and Punya Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes", *IEEE Proceedings* 1993, pp. 1064-1070.
- [16] R. Garelo, P. Pierleoni, and S. Benedetto, "Computing the free distance of Turbo Codes and Serially Concatenated Codes with Interleavers: Algorithms and Applications", *IEEE J. on Selected Areas in Comm.*, vol. 19, no. 5, pp. 800 – 812, May 2001.
- [17] M. Modarresi and S. Sadough, "Turbo-Trellis Coded Modulation Under Channel Estimation Inaccuracies", *IEEE Proceedings of ICEE* 2010, May 11-13, 2010.
- [18] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion in the presence of channel estimation errors, mobility, and multiple paths," *IEEE Transactions on Communications*, vol.47, no. 2, pp. 199-207, February 1999.

An Unsupervised Feature Selection Method Based On Genetic Algorithm

Nasrin Sheikhi, Amirmasoud Rahmani,
Mehran Mohsenzadeh

Department of computer engineering
Science and Research Branch, Islamic Azad University
khuzestan, Iran

Reza Veisisheikhrobat

National Iranian South Oil Company(NISOC)
Ahvaz, Iran

Abstract— In this paper we describe a new unsupervised feature selection method for text clustering. In this method we introduce a new kind of features that we called multi term features. Multi term feature is the combination of terms with different length. So we design a genetic algorithm to find the multi term features that have maximum discriminating power.

Keywords— multi term feature; discriminating power; genetic algorithm; fitness function

I. INTRODUCTION

Reducing dimensionality of a problem, in many real world problems, is an essential step before any analysis of the data. The general criterion for reducing the dimensionality is the desire to preserve most of the relevant information of the original data according to some optimality criteria. Dimensionality reduction or feature selection has been an active research area in pattern recognition, statistics and data mining communities. The main idea of feature selection is to choose a subset of input features by eliminating features with little or no predictive information. In particular, feature selection removes irrelevant features, increases efficiency of learning tasks, improves learning performance and enhances comprehensibility of learned results[2]

Depending on if the class label information is required, feature selection can be either unsupervised or supervised.

Feature selection has been well studied in supervised classification [3]. However, it is a quite recent research topic and also a challenging problem for clustering analysis for two reasons: first, it is not an easy task to define a good criterion for evaluating the quality of a candidate feature subset due to the absence of accurate labels of items. Second, it requires an exponentially increasing number of feature subset evaluations to optimize the defined criterion, that is in fact impractical if the data set has a large number of features.

Some methods for unsupervised feature selection have been proposed in the literature, such as document frequency(DF), term contribution(TC), Term Variance Quality(TVQ), Term Variance(TV) et al. In most of these methods a criterion is defined for evaluate the relevance of one term of documents for clustering, and depend on how much dimensionality reduction required, the number of most relevant features will be selected.

In this paper we proposed a novel feature selection method that evaluate the discriminating power of set of terms instead raw terms as features.

The main idea of this method is that a feature that is irrelevant by itself may become relevant when used with other features. So we describe new kind of feature named Multi Term Feature(MTF), that is the feature that made from combination of terms.

We use genetic algorithm for search the large space of different multi term features to find most relevant of them. To achieve this goal we designed the fitness function to estimate the discriminating power of MTFs.

The rest of this paper organized as follows: the next section describes two methods for evaluate relevance of MTFs. Section III explains using the genetic algorithm to find best MTFs. Experimental results are presented in section IV, and a conclusion is given in section V.

II. EVALUATE RELEVANCE OF MULTI TERM FEATURES

Because in many cases one term can not determine the subject of document very well, we use MTF to find the best terms that can determine the clusters of documents. So we must define criterions for evaluate relevance of MTFs. At first we must determine when a MTF appear in a document.

We defined appearance threshold for determine the presence of MTF in a document, that is the minimum number of terms of MTF that if appear in a document that's MTF appear in the document too.

Two criterions that we defined for evaluating discriminating power of MTFs are as follows:

A. Modified Term Variance

Term variance is one of the methods that use for evaluate the quality of term in dataset for clustering the documents. The equation of this method is as follows:

$$v(t_i) = \sum_{j=1}^N [f_{ij} - \bar{f}_i]^2 \quad (1)$$

In this method the terms that have high frequency but have not uniform distribution over document will have high TV value. We modified TV method to use with MTFs :

$$v(MTF_i, th) = \sum_{j=1}^N [vf_{ij,th} - \overline{vf_{i,th}}]^2 \quad (2)$$

In this relation $vf_{ij,th}$ is the frequency of i th MTF in document j with appearance threshold th , and $\overline{vf_{i,th}}$ is the average of i th MTF frequency in all documents. The frequency of MTF is measured by equation as follows:

$$vf_{ij,th} = \begin{cases} \frac{not}{length} * \sum_{k=1}^m f_{kj} & \text{if } d_j \text{ contains}(MTF_i, th) \\ 0 & \text{else} \end{cases} \quad (3)$$

In this relation f_{kj} is the k th term of MTF, m is the number of term of MTF and d_j is the j th document in dataset, not is the number of different MTF 's term that appear in d_j and length is the length of MTF.

$d_j \text{ contains}(MTF_i, th)$ is the logical function that determine if d_j contains the MTF return TRUE and else return FALSE.

$\overline{vf_{i,th}}$ is measured as follows:

$$\overline{vf_{i,th}} = \frac{\sum_{j=1}^N vf_{ij,th}}{N} \quad (4)$$

B. Dependency Between Terms

Another criterion that we define to evaluate the relevance of MTFs is dependency between terms of MTF that measure by this equation:

$$dep(MTF_i, th) = \frac{\sum_{j=1}^N vf_{ij,th}}{\sum_{j=1}^N vf_{ij,th=0}} \quad (5)$$

Our goal is to find the MTFs that have high discriminating power, so we look for find the MTFs that terms of them is belong to same subject and most of the time appear in the documents of that subject.

Dependency between the terms of MTF is the ratio of sum of MTF 's frequency in all documents to sum of the MTF 's terms frequency. This value show that most of the time the terms of MTF appear together in documents or separately.

III. USING GENETIC ALGORITHM

As we already mentioned our goal is to find best MTFs that can determine the clusters of documents. Because of the

large number of MTFs that can extract from documents, using a search algorithm that can search a huge amount of data is necessary.

Genetic algorithm is one of the best algorithms that can find best solutions for a problem between large number of solutions that is in the search space of problem. So we use genetic algorithm as our search strategy to find the most discriminating MTFs that exist in the search space.

A. Chromosomes

Each chromosome in this method is a MTF that can have different length. Each gene of chromosome is a term of the MTF. So the chromosome is shown as the set of terms and not a binary code.

B. Initial population

Initial population is the set of specific number of chromosomes.

C. Crossover and Mutation

The genetic algorithm generates new solutions by recombining the genes of the current best solutions. This is accomplished through the crossover and the mutation operators. On a one-point crossover, the crossing point is selected at random and genes from one side of the chromosomes are exchanged. In our model because of the different length of chromosomes crossover method is different too.

In this method the crossing point is selected at random on both of parent chromosomes. Then one side of chromosomes are exchanged, so two chromosomes of results of this kind of crossover have not equal length.

The mutation operator selected one position of gene in chromosome at random, and then exchange it with the term that is selected from documents randomly.

D. Fitness function

The objective function is the cornerstone of the genetic process. We designed the following fitness function to explore the space of solutions:

$$fitness(ch_i) = v(ch_i, th) * dep(ch_i, th) * \ln(length_{ch_i} + 1) \quad (6)$$

In this function:

ch_i Is the i th chromosome

$fitness(ch_i)$ Is the fitness value of ch_i

$v(ch_i, th)$ Is the modified term variance value of ch_i with appearance threshold th

$dep(ch_i, th)$ Is the value of dependency between terms of ch_i

$length_{ch_i}$ Is the length of ch_i

We described the modified term variance and dependency between terms of MTF in section

Another part of our fitness function is $\ln(\text{length}_{n_i} + 1)$

During the genetic process the chromosomes' length will increased because of operating crossover on the population. According to this length increasing the probability of presence of chromosome in the document and as its result the value of modified term variance and dependency criterions for this chromosome will decreased. So the fitness value of the chromosomes with more length will decreased, and then the algorithm go to select the smaller chromosome and so go to usual methods.

By adding $\ln(\text{length}_{n_i} + 1)$ we give more chance to the bigger chromosomes to be selected as relevance chromosome.

E. The proposed algorithm

The designed algorithm is as follows:

1. An initial set of solutions is established at random. This population contains chromosomes that are MTFs that made of terms that selected randomly from documents.
2. The fitness value of each chromosome is measured by fitness function, the stopping criteria are tested. As a general criterion the genetic process is stopped when the maximum fitness does not increase over a few iterations.
3. Selection, mutation and crossover operate on population.
4. The new population is generated and the iterative process buckles up from step 2.

IV. EXPERIMENTAL RESULTS

The following experiments we conducted are to compare the proposed genetic model and term variance method.

We choose 3 datasets from Reuters-21587 that each one have 1000 documents.

We choose K-means to be the clustering algorithm .since K-means clustering algorithm is easily influenced by selection of initial centroids, we random produced 10 sets of initial centroids for each dataset and averaged 10 times performance as the final clustering performance.

We use Average Accuracy(AA) and F1-Measure(F1) that defined in [3], as clustering validity criterions for evaluate the accuracy of clustering results. This results on reut2-001, reut2-002 and reut2-003 datasets are shown in Fig. 1 to Fig. 6.

From these figures, we can see that proposed algorithm can improve the clustering accuracy in most of experiments results.

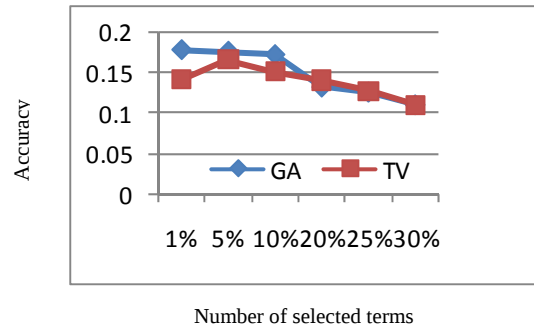


Figure 1. precision comparison on reut2-001(AA)

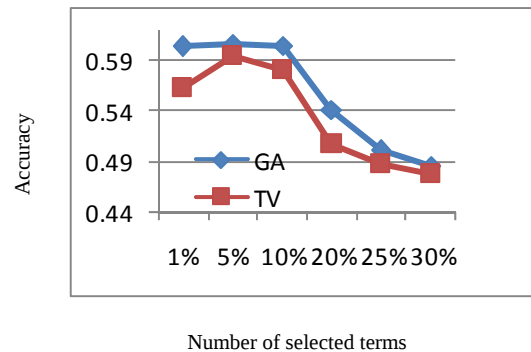


Figure 2. precision comparison on reut2-001(F1)

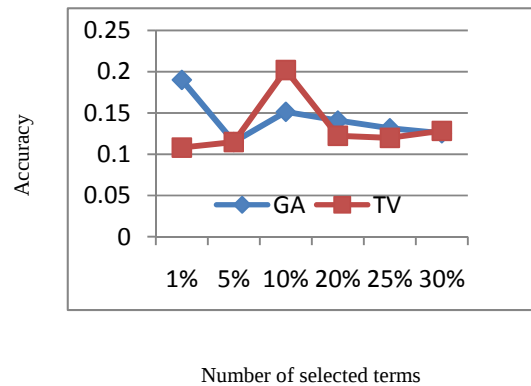


Figure 3. precision comparison on reut2-002(AA)

V. CONCLUSION

In this paper we described a new feature selection method based on genetic algorithm. We use the new kind of feature that we called MTF that is the set of terms and then define the criterions for evaluate the relevance of these features. The experimental results shown that the proposed method can improve the accuracy of clustering.

ACKNOWLEDGMENT

Athors thank national iranain oil company (nioc) and national iranain south oil company (nisoc) for their help and financial support.

REFERENCES

- [1] Sullivan, D., *Document warehousing and text Mining*, John Wiley, New York, 2001.J.
- [2] Miller, T., *Data and text mining a business applications approach*, Prentice Hall, New York, 2005.
- [3] Liu, L. and Kang, J. and YU, J. and Wang, Z., "A comparative study on unsupervised feature selection methods for text Clustering", *Proceeding of NLP-KE*, Vol. 9, pp. 597-601, 2005.
- [4] Aliane, H., "An ontology based approach to multilingual information retrieval", *IEEE Information and Communication Technologies*, Vol. 1, pp. 1732-1737, 2006.
- [5] Yu Lee, L. and Soo, v., "Ontology-based information retrieval and extraction", *International Conference on Information Technology: Research and Education*, Vol. 8, pp. 265-269, 2005.
- [6] Nayyeri, A. and Oroumchian, F., "FuFaIR: a fuzzy farsi information retrieval system", *IEEE International Conference on Computer Systems and Applications*, Vol. 3, pp. 1126-1130, 2006.
- [7] Desjardins, G. and Proulx, R. and Godin, R., "An auto-associative neural network for information retrieval", *IEEE International Joint Conference on Neural Networks*, Vol. 9, pp. 3492-3498, 2006.
- [8] Tian, Q., "A foundational perspective for visual information retrieval", *Multimedia IEEE*, Vol. 13, pp. 90-92, 2006.
- [9] Brunner, J. and Naudet, Y. and Latour, T., "Information retrieval in multimedia: exploiting MPEG-7 metadata by the use of ontologies and fuzzy thematic spaces", *Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05)*, Vol. 7, pp. 1-6, 2005.
- [10] Dong, A. and Li, H., "Multi-ontology based multimedia annotation for domain-specific information retrieval", *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06)*, Vol. 9, pp. 1-8, 2006.
- [11] Heo, S. and Motoyuki, S. and Ito, A. and Makino, S., "An effective music information retrieval method using three-dimensional continuous DP", *IEEE Transactions on Multimedia*, Vol. 8, NO. 3, pp. 633-639, 2006.
- [12] Chang, C. and Kayed, M., "A survey of web information extraction systems", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, NO. 10, pp. 1411-1428, 2006.
- [13] Kim, J. and Moldovan, D., "Acquisition of linguistic patterns for knowledge-based information extraction", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, NO. 5, pp. 713-724, 1995.
- [14] Ramshaw, A. and Weischeldel, M., "Information extraction", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 7, pp. 969-972, 2005.
- [15] Lam, M. and Gong, Z., "Web information extraction", *Proceedings of the 2005 IEEE International Conference on Information Acquisition*, Vol. 1, pp. 569-601, 2005.
- [16] Yang, S. and WU, X. and Deng, Z. and Zhang M. and Yang, D., "Relative term-frequency based feature selection for text

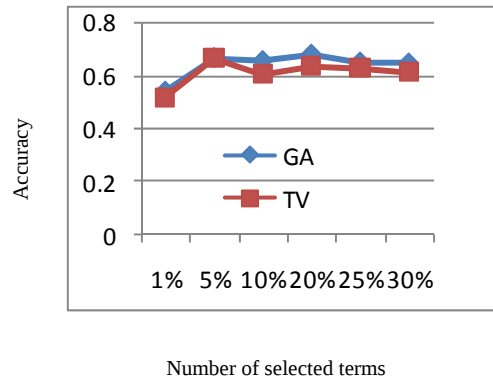


Figure 4. precision comparison on reut2-002(F1)

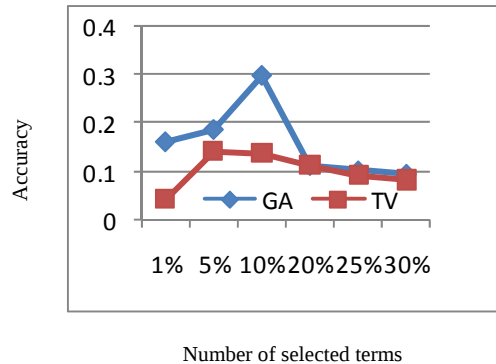


Figure 5. precision comparison on reut2-003(AA)

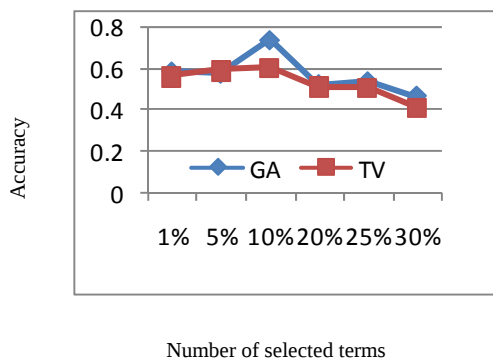


Figure 6. precision comparison on reut2-003(F1)

- categorization", *Proceeding of the IEEE first international Conference on Machine Learning and Cybernetics*, Vol. 4, pp. 1432-1436, 2002.
- [17] Yiming, Y. and Pedersen, J., "A comparative study on feature selection in text categorization", *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 412-420, 1997.
- [18] Prabowo, R. and Thelwall, M., "A comparison of feature selection methods for an evolving RSS feed corpus", *Information Processing and Management*, Vol. 42, pp. 1491-1512, 2006.
- [19] How, B. and Narayanan, K., "An empirical study of feature selection for text categorization based on Term weightage", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, Vol. 2, pp. 1-4, 2004.
- [20] Li, S. and Zong, C., "A new approach to feature selection for text categorization", *IEEE Proceeding of NLP-KE'05*, Vol. 9, pp. 626-630, 2005.
- [21] Mitchell, T., *Machine Learning*, McGraw-Hill, Washington, 1997.
- [22] Dong, Y. and Han, K., "A comparison of several ensemble methods for text categorization", *Proceedings of the 2004 IEEE International Conference on Services Computing (SCC'04)*, Vol. 4, pp. 1-4, 2004.
- [23] Soucy, P. and Mineau, G., "A simple KNN algorithm for text categorization", Vol. 8, pp. 647-648, 2001.
- [24] Namburu, S. and Tu, H. and Luo, J. and Pattipati, R., "Experiments on supervised learning algorithms for text categorization", *IEEE Aerospace Conference*, pp. 1-8, 2005.
- [25] Goldberg, J.L., "CDM: An approach to learning in text categorization", *Proceedings of Seventh International Conference on Tools with Artificial Intelligence*, Vol. 9, pp. 258-265, 1995.

A PCA Based Feature Extraction Approach for the Qualitative Assessment of Human Spermatozoa

V.S.Abbiramy

Department of Computer Applications
Velammal Engineering College
Chennai, India
vml_rithi@yahoo.co.in

Dr. V. Shanthi

Department of Computer Applications
St. Joseph's Engineering College
Chennai, India
drvshanthi@yahoo.co.in

Abstract— Feature extraction is often applied in machine learning to remove distracting variance from a large and complex dataset, so that downstream classifiers or regression estimators can perform better. The computational expense of subsequent data processing can be reduced with a lower dimensionality. Reducing data to two or three dimensions facilitates visualization and further analysis by domain experts. The data components (parameters) measured in computer assisted semen analysis have complicated correlations and their total number is also large. This paper presents Principal Component Analysis to de-correlate components, reduce dimensions and to extract relevant features for classification. It also compares the computation of Principal Component Analysis using the Covariance method and Correlation method. Covariance based Principal Component Analysis was found to be more efficient in reducing the dimensionality of the feature space.

Keywords- Covariance based PCA; Correlation based PCA; Dimensionality Reduction; Eigen Values; Eigen Vectors Feature Extraction; Principal Component Analysis.

I. INTRODUCTION

The term data mining refers to the analysis of large datasets. Humans often have difficulty comprehending data in many dimensions. Algorithms that operate on high-dimensional data tend to have a very high time complexity. Many machine learning algorithms and data mining techniques struggle with high-dimensional data. This has become known as the curse of dimensionality. Reducing data into fewer dimensions often makes analysis algorithms more efficient and can help machine learning algorithms make more accurate predictions. Thus Data reduction is considered as an important operation in the data preparation step.

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results. One of the strategies used for data reduction is dimensionality reduction. Dimensionality reduction reduces the data set size by removing attributes or dimensions which may be irrelevant to the mining task. The best and worst attributes are determined using tests of statistical significance which assumes that the attributes are independent of one another. Dimensionality reduction can be divided into feature selection and feature extraction.

Feature selection approaches try to find a subset of the original variables. Two strategies are filter (e.g. information

gain) and wrapper (e.g. search guided by the accuracy) approaches. Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions. The transformation technologies can be categorized into two groups: linear and non-linear methods. Linear methods use linear transforms (projections) for dimensional reduction, while the non-linear methods use non-linear transforms for the same purpose. The linear technologies include PCA, LDA, 2DPCA, 2DLDA and ICA. The non-linear dimensionality reduction technologies include KPCA and KFD.

II. RELATED WORK

Several linear and non linear methods have been discussed in the survey of dimension reduction techniques paper [1]. The paper reviews current linear dimensionality reduction techniques, such as PCA, Multidimensional scaling(MDS), and nonlinear dimensionality reduction techniques, such as Isometric Feature Mapping (Isomap), Locally Linear Embedding (LLE), Hessian Locally Linear Embedding (HLLE) and Local Tangent Space Alignment (LTSA). In order to apply nonlinear dimensionality reduction techniques effectively, the neighborhood, the density, and noise levels need to be taken into account [2]. A brief survey of dimensionality reduction methods for classification, data analysis and interactive visualization was given [10].

In paper [3], an effort has been made to predict the suitable time period within a year for mustard plant by considering the total effect of environmental parameters using the method of factor analysis and principal component analysis. The paper proposes a mechanism for comparing and evaluating the effectiveness of dimensionality reduction techniques in the visual exploration of text document archives. Multivariate visualization techniques and interactive visual exploration are studied [4]. The paper compares four different dimensionality reduction techniques, such as PCA, Independent Component Analysis (ICA), Random Mapping (RM) and statistical noise reduction algorithm and their performance are evaluated in the context of text retrieval [5].

The paper [6] examines a rough set Feature Selection technique which uses the information gathered from both the lower approximation dependency value and a distance metric which considers the number of objects in the boundary region and the distance of those objects from the lower approximation. The use of this measure in rough set feature selection can result

in smaller subset sizes than those obtained using the dependency function alone. The methods proposed in paper [7] is bi-level dimensionality reduction methods that integrate filter method and feature extraction method with the aim to improve the classification performance of the features selected. In level 1 of dimensionality reduction, features are selected based on mutual correlation and in level 2 selected features are used to extract features using PCA or LPP.

The paper presents an independent component analysis (ICA) approach to DR, to be called ICA-DR which uses mutual information as a criterion to measure data statistical independency that exceeds second-order statistics. As a result, the ICA-DR can capture information that cannot be retained or preserved by second-order statistics-based DR techniques [8]. In paper [9], KPCA is used has a preprocessing step to extract relevant feature for classification and to prevent from the Hughes phenomenon. Then the classification was done with a backpropagation neural network on real hyper spectral ROSIS data from urban area. Results were positively compared to the linear version (PCA).

The author has proposed a model and compared four dimensionality reduction techniques to reduce the feature space into an input space of much lower dimension for the neural network classifier. Among the four dimensionality reduction techniques proposed, Principal Component Analysis was found to be the most effective in reducing the dimensionality of the feature space [11]. In this study, a novel biomarker selection approach is proposed which combines singular value decomposition (SVD) and Monte Carlo strategy to early Ovarian Cancer detection. Comparative study and statistical analysis show that the proposed method outperforms SVM-RFE and T-test methods which are the typical supervised classification and differential expression detection based feature selection methods [12]. The application of three different dimension reduction techniques to the problem of classifying functions in object code form as being cryptographic in nature or not were compared. It is demonstrated that when discarding 90% of the measured dimensions, accuracy only suffers by 1% for this problem [13].

III. PRINCIPAL COMPONENT ANALYSIS

Often, the variables under study are highly correlated and they are effectively “saying the same thing”. It may be useful to transform the original set of variables to a new set of uncorrelated variables called principal components. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data [20].

The goal is to transform a given data set X of dimension M to an alternative data set Y of smaller dimension L . Equivalently, we are seeking to find the matrix Y , where Y is the Karhunen–Loève transform (KLT) of matrix X :

$$Y = KLT\{X\} \quad (1)$$

Algorithm for computing PCA [15] using the covariance method consists of the following steps

Step 1: Organize the data set

The microscopic examination of spermatozoa dataset comprises of M observations and each observation is described with L variables. This dataset is arranged as a set of N data vectors $X_1, \dots, X_n$ with each X_n representing a single grouped observation of the M variables.

Step 2: Calculate the empirical mean

Find the empirical mean of the previous step along each dimension $m = 1, \dots, M$ and place it into a mean vector u of dimensions $M \times 1$.

$$u[m] = \frac{1}{N} \sum_{n=1}^N X[m, n] \quad (2)$$

Step 3: Calculate the deviations from the mean

The input dataset is centered by subtracting the empirical mean vector u from each column of the data matrix X and it is stored in the $M \times N$ matrix B .

$$B = X - uh \quad (3)$$

where h is a $1 \times N$ row vector of all 1s.

Step 4: Find the covariance matrix

Calculate the $M \times M$ covariance matrix C from the outer product of matrix B with itself:

$$C = E[B \otimes B] = E[B \cdot B^*] = \frac{1}{N} \sum B \cdot B^* \quad (4)$$

where E is the expected value operator,

$\otimes$ is the outer product operator and

$*$ is the conjugate transpose operator.

Step 5: Find the eigenvectors and eigenvalues of the covariance matrix

Compute the matrix V of eigenvectors which diagonalizes the covariance matrix C :

$$V^{-1} C V = D \quad (5)$$

where D is the diagonal matrix of eigenvalues of C . Matrix D will take the form of an $M \times M$ diagonal matrix, where

$$D[p, q] = \lambda_m \quad \text{for } p = q = m \quad (6)$$

is the m th eigenvalue of the covariance matrix

Step 6: Rearrange the eigenvectors and eigenvalues

Sort the columns of the eigenvector matrix V and eigenvalue matrix D obtained in the previous step in the order of decreasing eigenvalue.

Step 7: Compute the cumulative energy content for each eigenvector

The cumulative energy content g for the m th eigenvector is the sum of the energy content across all of the eigenvalues from 1 through m :

$$g[m] = \sum_{q=1}^m D[q,q] \quad \text{for } m = 1, \dots, M \quad (7)$$

Step 8: Select a subset of the eigenvectors as basis vectors

Then save the first L columns of V as the $M \times L$ matrix W:

$$W[p,q] = V[p,q] \quad \text{for } p = 1, \dots, M \quad q = 1, \dots, L \quad (8)$$

where

$$1 \leq L \leq M$$

The vector g can be used as a guide to choose an appropriate value for L so that it above a certain threshold, like 90 percent. ie.

$$g[m = L] \geq 90\%$$

Step 9: Convert the source data to z-scores

Create an $M \times 1$ standard deviation vector s from the square root of each element along the main diagonal of the covariance matrix C:

$$s = \{s[m]\} = \sqrt{C[p,q]} \quad \text{for } p = q = m = 1 \dots M \quad (9)$$

Also calculate the $M \times N$ z-score matrix:

$$Z = \frac{B}{s.h} \quad (\text{Divide element-by-element}) \quad (10)$$

Step 10: Project the z-scores of the data onto the new basis

The projected vectors are the columns of the matrix

$$Y = W^* . Z = KLT \{X\} \quad (11)$$

where W^* is the conjugate transpose of the eigenvector matrix. The columns of matrix Y represent the Karhunen–Loeve transforms (KLT) of the data vectors in the columns of matrix X.

IV. EXPERIMENTAL RESULTS

TABLE I. SAMPLE DATA INPUT

Area	Perimeter	HeadLe n	Head Width	Eccen tricity	MidLen	TailLen	Orienta tion	Equiv Diam	Mean Dist	Mean Velocity	A	B	C	D
36.0000	46.7032	9.8914	5.3217	0.8429	15.6200	67.5700	78.5457	6.7703	147.7534	49.2500	1.00	0.00	0.00	0.00
158.000	191.2335	23.6182	8.6646	0.9303	19.1100	66.2300	81.9619	13.9116	121.5033	40.5000	1.00	0.00	0.00	0.00
1.0000	3.6280	1.1547	1.1547	0.9600	10.2300	44.6900	0.0000	1.1284	0.0000	0.0000	0.00	0.00	0.00	1.00
37.0000	25.4530	10.1335	5.3444	0.8496	19.1200	35.5400	88.9338	6.8637	17.0721	5.6900	0.00	1.00	0.00	0.00
10.0000	11.8454	4.2583	3.2083	0.6575	14.0000	74.3100	90.0000	3.5682	66.6333	22.2100	1.00	0.00	0.00	0.00
1.0000	3.6280	1.1547	1.1547	0.0000	15.0700	44.1600	0.0000	1.1284	20.0056	6.6685	0.00	1.00	0.00	0.00
7.0000	13.4457	5.7735	1.8145	0.9493	15.3500	38.7700	45.0000	2.9854	17.0721	5.6907	0.00	1.00	0.00	0.00
7.0000	10.1697	3.8791	2.4300	0.7795	15.3500	42.2700	45.0000	2.9854	66.6330	22.2110	1.00	0.00	0.00	0.00
7.6840	53.9280	18.7500	5.0551	0.9538	28.1250	68.7600	-89.9088	81.0472	53.8103	17.9367	0.00	1.00	0.00	0.00
4.5723	35.9100	1.6600	1.0698	0.9600	2.4900	45.5600	89.7594	87.2289	58.4698	19.4899	0.00	1.00	0.00	0.00
6.4139	48.7620	8.4622	5.2565	0.9597	12.6933	32.4500	-89.8078	84.2590	36.6666	12.2222	0.00	1.00	0.00	0.00
5.6677	42.7140	9.5653	5.9086	0.9599	14.3480	65.4700	-89.9634	85.3550	30.9075	10.3025	0.00	1.00	0.00	0.00
7.1918	52.1640	2.4652	1.0893	0.9596	3.6978	45.2600	89.8741	84.4477	7.6775	2.5592	0.00	0.00	1.00	0.00
8.1603	63.2520	8.8570	5.6096	0.9605	13.2855	35.8900	-89.9996	82.8648	110.0000	36.6700	1.00	0.00	0.00	0.00
6.5885	49.8960	9.9260	5.8168	0.9593	14.8890	45.3400	-89.6868	83.5839	46.0652	15.3500	0.00	1.00	0.00	0.00
3.9684	7.6860	7.6593	4.1154	0.9618	11.4890	30.9800	89.3761	89.8604	11.9580	3.9860	0.00	0.00	1.00	0.00

The input dataset as shown in Table I is constructed by combining the statistical measurement of morphological parameters of spermatozoon given in Table 4.1 and motility statistics given in Table 5 of paper [16] & [17].

This data set has fifteen features, so every sample is a fifteen dimensional vector. The covariance matrix and correlation matrix calculated from the dataset for the first five features are given in Table II and Table III.

TABLE II. COVARIANCE MATRIX FOR INPUT DATASET

	Area	Perimeter	Head Length	Head Width	Eccentrici ty
Area	1,483.97591	1,522.06560	171.88957	55.43896	0.97301
Perimeter	1,522.06560	2,017.41106	219.77496	72.61875	3.16497
Head Length	171.88957	219.77496	38.35266	12.39077	0.49326
Head Width	55.43896	72.61875	12.39077	5.24063	0.18160
Eccentri city	0.97301	3.16497	0.49326	0.18160	0.05923

TABLE III. CORRELATION MATRIX FOR INPUT DATASET

	Area	Perimeter	Head Length	Head Width	Eccentricity
Area	1.00000	0.87968	0.72051	0.62865	0.10378
Perimeter	0.87968	1.00000	0.79010	0.70625	0.28952
Head Length	0.72051	0.79010	1.00000	0.87399	0.32726
Head Width	0.62865	0.70625	0.87399	1.00000	0.32594
Eccentricity	0.10378	0.28952	0.32726	0.32594	1.00000

Based on the correlation matrix, eigen values are calculated. In the case of N independent variables, there are N eigen values. For the given dataset there are 15 eigen values. The proportion of total variance in input dataset explained by the i^{th} principal component is simply the ratio between the i^{th} eigen value and the sum of all eigen values. Cumulative proportion of variance is computed by adding the current and previous proportion of variance.

The Eigen values, proportion and cumulative proportion of variance for the Covariance matrix and Correlation matrix are shown in Table IV & Table V.

TABLE IV. EIGEN VALUES OF THE COVARIANCE MATRIX

Component	Eigen Value	Proportion	Cumulative Proportion
0	104,057.27016	0.50139	0.50139
1	63,852.65226	0.30767	0.80906
2	21,280.82120	0.10254	0.91160
3	15,102.20724	0.07277	0.98437
4	2,298.23956	0.01107	0.99545
5	714.88982	0.00344	0.99889
6	210.22987	0.00101	0.99990
7	10.57894	0.00005	0.99995
8	5.50902	0.00003	0.99998
9	3.38897	0.00002	1.00000
10	0.44942	0.00000	1.00000
11	0.23906	0.00000	1.00000
12	0.03046	0.00000	1.00000
13	0.00002	0.00000	1.00000
14	0.00000	0.00000	1.00000

TABLE V. EIGEN VALUES OF THE CORRELATION MATRIX

Component	Eigen Value	Proportion	Cumulative Proportion
0	83.20765	0.36981	0.36981
1	39.75628	0.17669	0.54651
2	30.98150	0.13770	0.68420
3	19.25778	0.08559	0.76979
4	18.04188	0.08019	0.84998
5	10.87774	0.04835	0.89832
6	9.87493	0.04389	0.94221
7	7.53003	0.03347	0.97568
8	2.52474	0.01122	0.98690
9	2.16730	0.00963	0.99653
10	0.71743	0.00319	0.99972
11	0.05215	0.00023	0.99995
12	0.01058	0.00005	1.00000
13	0.00000	0.00000	1.00000
14	0.00000	0.00000	1.00000

Table VI shows the index of the component and its equivalent proportion of variance contribution in terms of percentage for both the approaches.

TABLE VI. PRINCIPAL COMPONENT BASE

Index	Covariance based Proportion of variance	Correlation based Proportion of variance
0	50.139 %	36.981 %
1	30.767 %	17.669 %
2	10.254 %	13.770 %
3	7.277 %	8.559 %
4	1.107 %	8.019 %
5	0.344 %	4.835 %
6	0.101 %	4.389 %
7	0.005 %	3.347 %
8	0.003 %	1.122 %
9	0.002 %	0.963 %
10	0.000 %	0.319 %
11	0.000 %	0.023 %
12	0.000 %	0.005 %
13	0.000 %	0.000 %
14	0.000 %	0.000 %

A line graph is plotted with the data taken from Table VI and it looks like what is shown in the Fig 1. The x-axis or the horizontal axis shows the principal component and the y-axis or the vertical axis shows the cumulative proportion of variance.

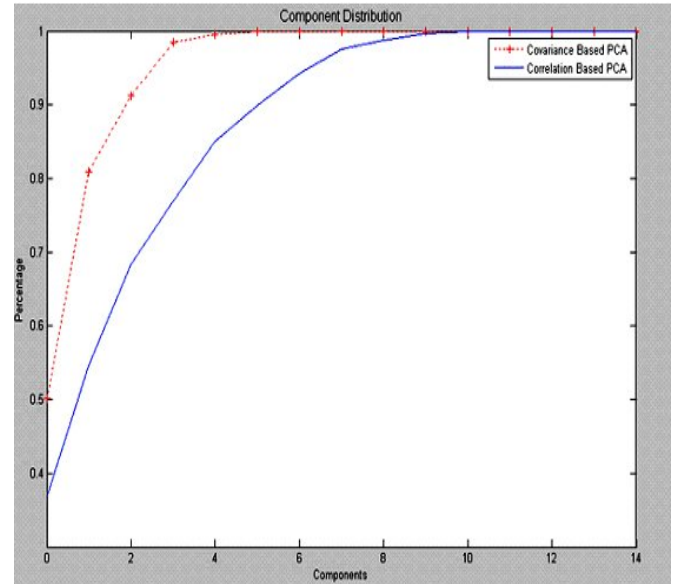


Figure 1. Visualization of Component Distribution For Covariance and Correlation Based PCA

A prior criterion has been set i.e. $g[m = L] \geq 90\%$ to select the number of principal components that will explain the maximal amount of variance. In the case of Covariance based PCA, the first 3 components explain 91.16% of proportion of variance and only these 3 components are extracted as shown in Table VII. In the case of Correlation based PCA, the first six components are extracted because it contributes to 89.83% of proportion of variance and their extraction is also shown in Table VIII.

TABLE VII. PROJECTION OF PRINCIPAL COMPONENTS FOR COVARIANCE BASED PCA

Covariance Based PCA		
Compt 1	Compt 2	Compt 3
84.0616	-55.2579	-51.1463
118.3525	-190.921	18.4082
-13.8898	61.9433	-33.3558
77.4735	39.0532	-2.522
79.0187	28.8897	-33.092
-11.803	50.2057	-41.2432
30.7564	52.1994	-22.8442
36.0141	25.2633	-43.0882
-108.986	-31.1044	1.971
53.9394	20.8128	49.9273
-112.899	-12.4223	13.7731
-113.766	-8.3706	11.6755
50.9376	38.7547	72.207
-103.324	-65.5479	-11.0093
-111.191	-20.1258	8.7022
45.3036	66.6282	61.6368

TABLE VIII. PROJECTION OF PRINCIPAL COMPONENTS FOR
CORRELATION BASED PCA

Correlation Based PCA					
Compt 1	Compt 2	Compt 3	Compt 4	Compt 5	Compt 6
3.2070	-2.0161	-0.3365	1.0866	-0.2185	0.3222
6.3144	-0.4150	0.9929	-2.0790	0.5209	-0.7784
-2.7463	-1.4483	-0.5652	-1.9583	-3.1207	-0.0414
-0.4117	0.6218	-0.7872	-1.5193	1.0358	-0.6042
0.7179	-2.3657	-1.0351	0.5842	0.2187	1.1480
-2.1965	-0.7410	-2.8671	0.1273	1.8017	-0.1998
-1.5856	0.1583	-0.8602	-0.6454	0.4515	-0.5720
0.0840	-1.9251	-0.6756	0.7497	-0.1392	0.0548
1.4450	2.8434	-1.0612	-0.2360	-0.1845	1.7155
-1.5676	-0.3117	0.8458	1.1044	0.2757	-1.5340
-0.6453	2.1079	0.2244	0.5307	-0.3289	-0.9156
-0.0133	2.1829	-0.2519	0.3425	-0.3734	0.5780
-2.4107	-0.9544	2.9752	-0.2137	0.7982	0.5641
1.7370	0.1307	0.7736	2.0326	-1.2414	-0.4710
-0.0076	2.1991	-0.0271	0.5050	-0.3525	-0.2792
-1.9204	-0.0668	2.6550	-0.4115	0.8565	1.0133

Table IX shows the comparison on the number of principal components extracted based on various criterions like

Kaiser Criterion: The eigen value is compared to 1 and only components with eigen values higher than 1 are retained.

Percentage of variance criterion: The number of components to be extracted is determined by the total explained percentage of variance.

Scree test: The eigen values are plotted against the principal components in a simple line plot as shown in Fig 1. It suggests finding the place where the smooth decrease of eigen values appears to level off to the right of the plot.

TABLE IX. COMPARISON ON NUMBER OF PRINCIPAL
COMPONENTS EXTRACTED

	Number of Dimensions in Input Dataset	No. of Components Extracted		
		Kaiser criterion	Percentage of variance>90%	Scree Test
Covariance Based PCA	15	10	3	4
Correlation Based PCA	15	10	6	8

V. CONCLUSION

Thus principal component analysis aims to summarize the data with many independent variables to a smaller set of derived variables in such a way that first component has maximum variance, followed by second and so on. This technique is particularly useful when a data reduction procedure is required that makes no assumptions concerning an underlying casual structure responsible for covariance in the data.

In this paper, we have compared the computation of Principal Component Analysis using the Covariance method and Correlation method and their performance is evaluated in terms of the number of components to be extracted. It was observed that Covariance based Principal Component Analysis

was found to be more efficient in reducing the dimensionality of the feature space.

REFERENCES

- [1] Imola K. Fodor, "A survey of dimension reduction techniques Center for Applied Scientific Computing", Lawrence Livermore National Laboratory, June 2002.
- [2] Flora S. Tsai and Kap Luk Chan, "Dimensionality Reduction Techniques for Data Exploration, Nanyang Technological University, ICICS 2007, 1-4244-0983-7/07/\$25.00 ©2007 IEEE
- [3] Satyendra Nath Mandal, J.Pal Choudhury ,S.R.Bhadra Chaudhuri,Dilip De, " Application of PCA & Factor Analysis for Predicting Suitable Period for Mustard Plant", International Conference on Computer and Electrical Engineering, IEEE Computer Society, 978-0-7695-3504-3/08 \$25.00 © 2008 IEEE
- [4] Shipping Huang, Matthew O. Ward and Elke A. Rundensteiner, "Exploration of Dimensionality Reduction for Text Visualization", Proceedings of the Third International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'05), IEEE Computer Society, 0-7695-2396-X/05 \$20.00 © 2005 IEEE
- [5] Vishwa Vinay, Ingemar J. Cox Ken Wood, Natasa Milic-Frayling,"A Comparison of Dimensionality Reduction Techniques for Text Retrieval", Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05), IEEE Computer Society, 0-7695-2495-8/05 \$20.00 © 2005 IEEE
- [6] Neil Mac Parthalain, Qiang Shen, and Richard Jensen, "A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction", IEEE Transactions on Knowledge And Data Engineering, Vol. 22, No. 3, March 2010
- [7] Veerabhadrapa, Lalitha Rangarajan, "Bi-level dimensionality reduction methods using feature selection and feature extraction", International Journal of Computer Applications (0975-8887) Volume 4 – No.2, July 2010
- [8] Jing Wang, Chein-I Chang, "Independent Component Analysis-Based Dimensionality Reduction With Applications in Hyperspectral Image Analysis", IEEE Transactions On Geoscience And Remote Sensing, Vol. 44, No. 6, June 2006
- [9] Mathieu Fauvel, Jocelyn Chanussot, Jon Atli Benediktsson, "Kernel Principal Component Analysis For Feature Reduction In Hyperspectral Images Analysis", University of Iceland and the Jules Verne Program of the French and Icelandic governments (PAI EGIDE).
- [10] Andreas Konig, "Dimensionality Reduction Techniques for Multivariate Data Classification, Interactive Visualization, and Analysis - Systematic Feature Selection vs. Extraction", Fourth International Conference on knowledge-Based Intelligent Engineering Systems & Allied Technologies, 30th Aug-1st Sep 2000. Brighton,UK.
- [11] Savio L.Y. Lam & Dik Lun Lee, "Feature Reduction for Neural Network Based Text Categorization", Department of Computer Science, Hong Kong University of Science and Technology
- [12] Shufei Chen, Bin Han, Lihua Li, Lei Zhu, Haifeng Lai, Qi Dai, "SVD Based Monte Carlo Approach to Feature Selection for Early Ovarian Cancer Detection", Institute for Biomedical Engineering and Instrumentation, Hangzhou Dianzi University
- [13] Jason L. Wright and Milos Manic, "The Analysis of Dimensionality Reduction Techniques in Cryptographic Object Code Classification", Idaho National Laboratory, University of Idaho, USA, Rzeszow, Poland, May 13-15, 2010.
- [14] Jiawei Han, Micheline Kamber, "Data mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, pp. 77-80.
- [15] http://en.wikipedia.org/wiki/Feature_extraction
- [16] V.S.Abbiramy , Dr. V. Shanthi, "Spermatozoa Segmentation and Morphological Parameter Analysis Based Detection of Teratozoospermia", International Journal of Computer Applications (0975 – 8887), Volume 3 – No. 7, June 2010
- [17] Ms. V.S.Abbiramy, Dr.V.Shanthi, Ms. Charanya Allidurai , "Spermatozoa Detection, Counting and Tracking in Video Streams to

Detect Asthenozoospermia.”, Internaional Conference on Signal and Image Processing, December 2010.

- [18] Mehmed Kantardize, “Data Mining: Concepts, Models, Methods, and Algorithms”, John Wiley & Sons © 2003, ISBN: 9780471228523.
- [19] Oded Maimon and Lior Rokach, “Data Mining and Knowledge Discovery Handbook”, Tel-Aviv University, Israel, ISBN-10: 0-387-24435-2
- [20] Ranjana Agarwal, A. R. Rao,”Data Reduction Techniques”,
“http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%20files/Manual%20II/9-data_reduction.pdf”

AUTHORS PROFILE

Ms. V S Abbiramy received the Master of Computer Applications from Sri Sarada Niketan college of Science for Women, Bharathidasan University, Tamil Nadu, India in 1998 and received the Master of Philosophy in Computer Science from Madurai Kamaraj University, Madurai in 2007. Currently she is working as Senior Lecturer, in Department of Computer Applications, Velammal Engineering College, Anna University, Chennai She is doing part time research in Mother Teresa University, Tamil Nadu, India. Her areas of interests include Medical Data Mining, Image Processing and Neural Networks. She has published 2 papers in National Conferences and 1 paper in International Conferences. She has published her work in 1 international journal and 2 in National level journal

Dr. V. Shanthi, M.Sc., M.Phil., PhD is working as a Professor in the Department of Computer Applications in St. Joseph’s College of Engineering , Chennai affiliated to Anna University. She has published one book and published 7 research papers in National Journals and 8 in International Journals and 14 Conferences.

Analysis of Error Metrics of Different Levels of Compression on Modified Haar Wavelet Transform

¹ T.Arumuga Maria Devi, Assistant Professor, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli .TamilNadu.

² S.S.Vinsley, Student IEEE Member, Guest Lecturer, Manonmaniam Sundaranar University, Centre for Information Technology and Engg, Manonmaniam Sundaranar University, Tirunelveli. TamilNadu.

Abstract—Lossy compression techniques are more efficient in terms of storage and transmission needs. In the case of Lossy compression, image characteristics are usually preserved in the coefficients of the domain space in which the original image is transformed. For transforming the original image, a simple but efficient wavelet transform used for image compression is called Haar Wavelet Transform. The goal of this paper is to achieve high compression ratio in images using 2 D Haar Wavelet Transform by applying different compression thresholds for the wavelet coefficients and these results are obtained in fraction of seconds and thus to improve the quality of the reconstructed image i.e., to arrive at an approximation of our original image. Another approach for lossy compression is, instead of transforming the whole image, to separately apply the same transformation to the regions of interest in which the image could be divided according to a predetermined characteristic. The Objective of the paper deals to get the coefficients is nearly closer to zero. More specifically, the aim of the thesis is to exploit the correlation characteristics of the wavelet coefficients as well as second order characteristics of images in the design of improved lossy compression systems for medical images. Here a modified simple but efficient calculation schema for Haar Wavelet Transform.

Index Terms—Haar Wavelet Transform – Linear Algebra Technique – Lossy Compression Technique – MRI

I. INTRODUCTION

IN recent years, many studies have been made on wavelets. An excellent overview of what wavelet have brought to the fields as diverse as biomedical applications, wireless communication, computer graphics or turbulence [30]. Image compression is one of the most visible applications of wavelets. The rapid increase in the range and use of electronic imaging justifies attention of systematic design of an image compression system and for providing the image quality needed in different applications. A typical still image

contains a large amount of spatial redundancy in plain areas where adjacent picture elements (pixels, pels) have almost same values. It means that the pixel values are highly correlated [31]. The basic measure for the performance of a compression algorithm is Compression Ratio (CR). In a lossy compression scheme, the image compression algorithm should achieve a tradeoff between compression ratio and image quality [32]. The balance of the paper is organized as follows: In section II describes the properties and advantages of haar wavelet transformation. In section III considers the procedure for haar wavelet transformation. In section IV, Implementation methodologies for wavelet compression and linear algebra are discussed. In section V describes the algorithm for its implementation. In section VI considers the comparison for metrics. In section VII considers the graphs, HWT image compression output results, and these results are plotted between various parameters. In section VIII elaborates on the importance of this paper, some applications and its extensions. Quality and compression can also vary according to input image characteristics and content. Images need not be reproduced exactly. An approximation of the original image is enough for most purposes, as long as the error between the original and the compressed image is tolerable. Lossy compression technique can be used in this area.

II. PROPERTIES AND ADVANTAGES OF HAAR WAVELET TRANSFORM

The Properties of the Haar Transform are described as follows:

- Haar Transform is real and orthogonal. Therefore
$$Hr = Hr^* \quad (1)$$
$$Hr^{-1} = Hr^T \quad (2)$$
Haar Transform is a very fast transform .

- The basis vectors of the Haar matrix are sequentially ordered.
- Haar Transform has poor energy compaction for images.
- Orthogonality: The original signal is split into a low and a high frequency part, and filters enabling the splitting without duplicating information are said to be orthogonal.
- Linear Phase: To obtain linear phase, symmetric filters would have to be used.
- Compact support: The magnitude response of the filter should be exactly zero outside the frequency range covered by the transform. If this property is satisfied, the transform is energy invariant.
- Perfect reconstruction: If the input signal is transformed and inversely transformed using a set of weighted basis functions, and the reproduced sample values are identical to those of the input signal, the transform is said to have the perfect reconstruction property. If, in addition, no information redundancy is present in the sampled signal, the wavelet transform is, as stated above, orthonormal.

No wavelets can possess all these properties, so the choice of the wavelet is decided based on the consideration of which of the above points are important for a particular application. Haar-wavelet, Daubechies-wavelets and biorthogonal-wavelets are popular choices [1]. These wavelets have properties which cover the requirements for a range of applications.

The advantages of Haar Wavelet transform are as follows:

1. Best performance in terms of computation time.
2. Computation speed is high.
3. Simplicity
4. HWT is efficient compression method.
5. It is memory efficient, since it can be calculated in place without a temporary array.

III. PROCEDURE FOR HAAR WAVELET TRANSFORM

To calculate the Haar transform of an array of n samples:

1. Find the average of each pair of samples. ($n/2$ averages)
2. Find the difference between each average and the samples it was calculated from. ($n/2$ differences)
3. Fill the first half of the array with averages.
4. Fill the second half of the array with differences.
5. Repeat the process on the first half of the array. (The array length should be a power of two)

IV. IMPLEMENTATION METHODOLOGY

Each image [27] is presented mathematically by a matrix of numbers. Haar wavelet uses a method for manipulating the matrices called averaging and differencing. Entire row of a image matrix is taken, then do the averaging and differencing process. After we treated entire each row of an image matrix, then do the averaging and differencing process for the entire each column of the image matrix. Then consider this matrix is known as a semifinal matrix (T) whose rows and columns have been treated. This procedure is called wavelet transform.

Then compare the original matrix and last matrix that is semifinal matrix(T), the data has become smaller. Since the data has become smaller, it is easy to transform and store the information. The important one is that the treated information is reversible. To explain the reversing process we need linear algebra. Using linear algebra is to maximize compression while maintaining a suitable level of detail.

A. Wavelet Compression Methodology

From Semi final Matrix (T) is ready to be compressed. [29] Definition of Wavelet Compression is fix a non negative threshold value ϵ and decree that any detail coefficient in the wavelet transformed data whose magnitude is less than or equal to zero (this leads to a relatively sparse matrix). Then rebuild an approximation of the original data using this doctored version of the wavelet transformed data. In the case of image data, we can throw out a sizable proportion of the detail coefficients in this and obtain visually acceptable results. This process is called lossless compression. When no information is loss (eg., if $\epsilon=0$). Otherwise it is referred to as lossy compression (in which case $\epsilon > 0$). In the former case, we can get our original data back, and in the latter we can build an approximation of it. Because we know this, we can eliminate some information from our matrix and still be capable of attaining a fairly good approximation of our original matrix. Doing this, we take threshold value $\epsilon=10$ ie., reset to zero all elements of semifinal matrix(T) which are less than or equal to 10 in absolute value. From this we obtain the doctored matrix. Then apply the inverse wavelet transform to doctored matrix we get the reconstructed approximation R.

In this process, we can get a good approximation of the original image. We have lost some of the detail in the image but it is so minimal that the loss would not be noticeable in most cases.

B. Linear Algebra Methodology

To apply the averaging and differencing using linear algebra [27] we can use matrices such as $A_1, A_2, A_3, \dots, A_n$ that perform each of the steps of the averaging and differencing process.

- i. When multiplying the string by the first matrix of the first half of columns are taking the average of each pair and the last half of columns take the corresponding differences.
- ii. The second matrix works in much the same way, the first half of columns now perform the averaging and differencing to the remaining pairs, and the identity matrix in the last half of columns carry down the detail coefficient from step i.
- iii. Similarly in the final step, the averaging and differencing is done by the first two columns of the matrix, and the identity matrix carries down the detail coefficient from previous step.
- iv. To simplify this process, we can multiply these matrices together to obtain a single transform matrix $W=A_1A_2A_3$ we can now multiply our original string by just one transform matrix to go directly from the original string to the final results of step iii.
- v. In the following equation we simplify this process of matrix multiplication. First the averaging and differencing and second the inverse of those operation.

$$\begin{aligned} 1. \quad T &= ((AW)^T W)^T \\ T &= (W^T A^T W)^T \\ T &= W^T (A^T)^T (W^T)^T \\ T &= W^T A W \end{aligned} \quad (3)$$

$$\begin{aligned} 2. \quad (W^1)^{-1} T W^{-1} &= A \\ (W^{-1})^1 T W^{-1} &= A \end{aligned} \quad (4)$$

V. ALGORITHM

- Read the image from the user.
- Apply 2 D DWT using haar wavelet over the image
- For the computation of haar wavelet transform, set the threshold value 25%, 10%, 5%, 1% ie., set all the coefficients to zero except for the largest in magnitude 25%, 10%, 5%, 1% . And reconstruct an approximation to the original image by apply the corresponding inverse transform with only modified approximation coefficients.
- This simulates the process of compressing by factors of $\frac{1}{4}, \frac{1}{10}, \frac{1}{20}, \frac{1}{100}$ respectively.
- Display the resulting images and comment on the quality of the images.
- Calculate MSE, MAE & PSNR values of different Compression Ratios for corresponding Reconstructed images.
- Then add a small amount of white noise to the input image. Default : variance =0.01, sigma=0.1, mean=0
- To compute the haar wavelet transform, set all the approximation coefficients to zero except those whose magnitude is larger than 3 sigma.
- This same case is applicable to detail coefficients that is horizontal, vertical & diagonal coefficients.
- Reconstruct an estimate of the original image by applying the corresponding inverse transform.
- Display and compare the results by computing the root mean square error, PSNR, and mean absolute error of the noisy image and the denoising image.
- The same process is repeated for various images and compare its performance.

Alternative approach Algorithm is described as follows:

1. Read the image cameraman.tif from the user.
2. Using 2D wavelet decomposition with respect to a haar wavelet computes the approximation coefficients matrix CA and detail coefficient matrixes CH, CV, CD (horizontal, vertical & diagonal respectively) which is obtained by wavelet decomposition of the input matrix ie., im_input.
3. From this, again using 2D wavelet decomposition with respect to a haar wavelet computes the approximation and detail coefficients which are obtained by wavelet decomposition of the CA matrix. This is considered as level 2.
4. Again apply the haar wavelet transform from CA matrix which is considered as CA1 for level 3.
5. Do the same process and considered as CA2 for level 4
6. Take inverse transform for level 1, level 2, level 3 & level 4 that ie., im_input, CA, CA1, CA2.
7. Reconstruct the images for level 1, level 2, level 3 & level 4.
8. Display the results of reconstruction 1, reconstruction 2, reconstruction 3, reconstruction 4 ie., level 1 , 2 , 3 , 4 with respect to the original image.

VI. METRICS FOR COMPARISON

A. Compression Metrics

The type of compression metrics [21] used for digital data varies quite markedly depending on the type of compression employed that use a simple ratio formula as given below

$$\text{CompressionRatio(\%)} = \frac{\text{Output File size (bytes)}}{\text{Input File size (bytes)}} \times 100 \quad (5)$$

A measure of rate is far more suitable metric for large compression ratios as it gives the number of bits per pixel used to encode an image rather than some abstract percentage.

$$\text{Rate(bpp)} = \frac{8 \times \text{Output File size (bytes)}}{\text{Input File size (bytes)}} \quad (6)$$

Rate, for image coding purposes, uses bits per pixel (bpp) as its unit.

B. Error Metrics

In general, error measurements are used on lossy compressed images to try and quantify the quality of a picture. Getting a quantifiable measure of the distortion between two images is very important as one can try and minimize this thesis so as to better replicate the original image. There are many ways of measuring the fidelity of a picture $g(x,y)$ to its original $f(x,y)$. One of the simplest and most popular methods is to use the difference between f and g . In its most basic form is the mean square error (MSE) [11] given by,

$$\text{MSE} = \frac{1}{N_1 N_2} \sum_{y=1}^{N_1} \sum_{z=1}^{N_2} (f(x,y) - g(x,y))^2 \quad (7)$$

where f and g are $N_1 \times N_2$ size image. This is a very useful measure as it gives an average value of the energy loss in the lossy compression of the original image f . Signal-to-noise ratio (SNR) is another measure often used to compare the performance of reproduced images which is defined by,

$$\text{SNR} = 10 \times \log_{10} \left[\frac{\frac{1}{N_1 N_2} \sum_{y=1}^{N_1} \sum_{z=1}^{N_2} f(x,y)^2}{\frac{1}{N_1 N_2} \sum_{y=1}^{N_1} \sum_{z=1}^{N_2} (f(x,y) - g(x,y))^2} \right] \quad (8)$$

SNR is measured in dB's and gives a good indication of the ratio of signal to noise reproduction. This is a very similar measurement to MSE. A more subjective qualitative measurement of distortion is the Peak Signal-to-noise ratio (PSNR) [3].

$$\text{PSNR} = 10 \times \log_{10} \left[\frac{255^2}{\frac{1}{N_1 N_2} \sum_{y=1}^{N_1} \sum_{z=1}^{N_2} (f(x,y) - g(x,y))^2} \right] \quad (9)$$

And also Mean Absolute Error is denoted by

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |g(x,y) - f(x,y)| \quad (10)$$

The standard error measures a large distortion, but the image has merely been brightened. Individually, MSE, SNR and PSNR are not very good at measuring subjective image quality, but used together these error metrics are at least adequate at determining if an image is reproduced at a certain quality.

VII. RESULTS AND DISCUSSIONS

The project deals with the implementation of the haar wavelet compression techniques and a comparison over various input images. We first look in to results of wavelet compression technique by calculating their comparison ratios and then compare their results based on the error metrics which is shown in Table I.

Table I
Different types of Error Metrics with respect to Various Compression Ratios.

Compression Ratios	Error Metrics		
	MSE	RMSE	PSNR
4:01	9937.92	99.6891	18.7841
10:01	14380.1	119.917	15.0893
20:01	15990.1	126.452	14.028
100:1	17453.28	132.1109	13.1524

A. Effects on Compression Ratio Vs Various Parameters

Wavelet Compression is applied for all images and the compression ratio is being calculated. The image quality thus measured in compression techniques is compared using a BAR CHART, which proves the image quality of the various input images which reconstructed images are better. This is shown in Fig 1 & 2.

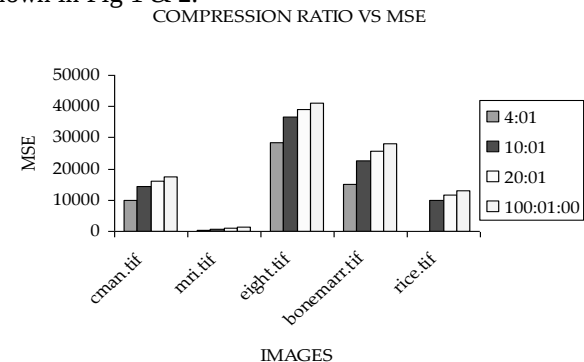


Fig. 1. Effects on Compression Ratio Vs MSE with Respect to Various Input Images

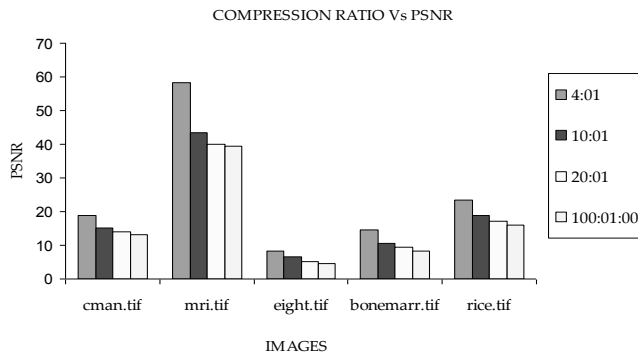


Fig. 2. Effects on Compression Ratio Vs PSNR with Respect to various input images

In the first experiment, we report on Compressed Ratio consumed by Error Metrics as described in the previous section. In our experiments, we used the gray scale sample, cameraman.tif of size 256*256. And measured the compression ratio and the PSNR of the compressed image. In each case, the threshold level was changed and got the output. The results are presented in Fig 5. The x-axis represents Compression Ratio, while the y-axis represents MSE with various input images.

In the Second experiment, we report on Compression Ratio Vs PSNR with various input images. In our experiments, we used the gray scale image samples. In each case, the threshold level was changed and got the output. The results are presented in Fig 5. The x-axis represents Compression ratio, while the y-axis represents PSNR with various input images.

The quality of compressed image depends on the no of decompositions. The no of decompositions determines the resolution of the lowest level in wavelet domain. Using larger no of decompositions, that will be more successful in resolving important HWT coefficients from less important coefficients. After decomposing the image and representing it with wavelet coefficients, compression can be performed by ignoring all coefficients below some threshold. In this experiment, compression is obtained by wavelet coefficient thresholding. All coefficients below some threshold are neglected and Compression Ratio is computed. Compression Algorithm operation is follows : Compression Ratio is fixed to the required level and threshold value has been changed to achieve required Compression Ratio after that PSNR is computed. PSNR tends to saturate for a larger no of decompositions. For each compression ratio, the PSNR characteristic has “threshold” which represents the optimal no of decompositions. Below and above the threshold PSNR decreases and no of decomposition increases. PSNR is increased up to some no of decompositions. Beyond that, increasing the no of decomposition has a negative effect.

PSNR values depend on image type and cannot be used to compare images with different content.

The quality of the reconstructed image is measured using the error metrics:

- MSE
- PSNR

The values of MSE & PSNR are calculated for all test images and are provided in Table II.

Table II

Calculation of different kinds of Error Metrics with respect to Various Input images under Compression Ratio 100:1

SI.No	Image	No of elements to store in bytes	MSE	PSNR
1	cman.tif	256*256	17543.280	13.1524
2	Rice.tif	256*256	13094.653	16.0257
3	Mri.tif	128*128	1248.827	39.5257
4	Eight.tif	242*308	41227.182	4.5567
5	Bonemarr.tif	238*270	28009.786	8.4222

VIII. CONCLUSION

This paper reported is aimed at developing computationally efficient and effective algorithm for lossy image compression using wavelet techniques. This paper is particularly targeted towards wavelet image compression using Haar Transformation with an idea to minimize the computational requirements to achieve good reproduction image quality. In this direction the following methods are developed. This image compression schemes for images have been presented based on the 2 D HWT. The promising results obtained concerning reconstructed image quality as well as preservation of significant image details, while on the otherhand achieving high compression rates.

- High compression ratio and better image quality obtained which is better than existing methods.
- In addition, the above methods are to be for noisy images
- To improve the quality of the reconstructed image
- The results are executed in fraction of seconds.
- To obtain the wavelet coefficients are nearly closer to zero.
- Image denoising techniques will allow precise imaging at much faster rates by greatly reducing the necessary averaging time to construct low noise images.

The performance of an image compression algorithm is basically evaluated in terms of compression ratio, MSE, and PSNR. A 'good' algorithm has a high compression ratio. Wavelet based image compression theory has rapidly increased in the last seven years. With the increasing use of multimedia technologies image compression requires higher performance as well as new functionality. To address this need in the specific area of still image compression here a new compression technique is proposed. The results proved that the compression ratio is very high and the reconstruction is same as that of the original image.

Wavelet based image compression indeed is a new and emerging area and has a lot of scope for improvement and extension. In future the following aspects may be considered for improving the algorithm.

This paper has focused on development of efficient and effective algorithm for still image compression. Fast and lossy coding algorithm using wavelet is developed.

Results shows that reduction in encoding time with little degradation in image quality compared to existing methods. While comparing the developed method with other methods compression ratio is also increased.

Some of the applications require a fast image compression technique but most of the existing technique requires considerable time. So this proposed algorithm developed to compress the image so fastly.

Wavelet transform is popular in image compression mainly because of its multi resolution and high energy compaction properties. Pictures or images are non stationary in frequency and spatial content; hence, data representing picture content may be any where in the actual picture. This property is included in this thesis.

The main bottleneck in the compression lies in the search of domain, which is inherently time expensive. This leads to excessive compression time. The algorithm can be improved by applying some indexing scheme.

The algorithm needs to be explored and tuned for domain specific application one of the major application area lies in the satellite imagery. As large quantity remote sensing data are collected on a regular basis for different level resource monitoring, there is an increasing need for compression for better data management. Through investigations of different categories of satellite imagery are required and SNR, depth of search etc. are to be studied to determine the usability of fractal image compression.

Better visual modules and perception based error criteria are needed for image coding. Using wavelet number of

implementation issues such as bit allocation methods and error estimation can be studied.

Image denoising method using wavelet for noisy image could be developed. This yield better result in image compression techniques using wavelet for noisy input images.

IX. REFERENCES

- [1] Sonja Grgic, Mislav Grgic, Member, IEEE, and Branka Zovko-Cihlar, Member, IEEE, "Performance Analysis of Image Compression Using Wavelets", IEEE Trans. Vol. 48., No 3, 2001.
- [2] Detlev Marpe, Member, IEEE, Gabi Blättermann, Jens Rieke, and Peter Maa, "A Two-Layered Wavelet-Based Algorithm for Efficient Lossless and Lossy Image Compression", IEEE transactions on circuits and systems for video technology, vol. 10, no. 7, october 2000.
- [3] Rafael C Gonzalez, Richard E Woods "Digital Image Processing" Pearson Asia Education 2nd edition.
- [4] Anil K Jain "Fundamental of image processing" Pearson Asia Education.
- [5] Mulcahy, colm PH.D, " Image Compression using haar wavelet transform", Spelman Science and Math Journal.
- [6] Corban Radu, Ungureanu Iulian, "DCT Transform and Wavelet Transform in Image Compression Applications".
- [7] Amara Graps, "An Introduction to Wavelets," IEEE Computational Science and Engineering, vol. 2, no. 2, Summer 1995.
- [8] Wim Sweldens and Peter Schroder, "Building Your Own Wavelet at Home".
- [9] Christopher Torrence and Gilbert P. Compo Program in Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado, "A Practical Guide to Wavelet Analysis".
- [10] Martin Vetterl, "Wavelets, Approximation, and Compression" 2001 IEEE Signal Processing Magazine.
- [11] Bryan E. Usevitch "A Tutorial on modern lossy wavelet Image Compression Foundation of JPEG 2000", IEEE signal processing magazine.
- [12] D.A. Karras, S.A. Karkanis and B.G. Mertzios, "Image Compression Using the Wavelet Transform on Textural Regions of Interest".
- [13] Ligang Ke And Micheal W. Marcellin, "Near Lossless Image Compression: Minimum Entropy, Constrained-Error DPCM", Department Of Electrical & Computer Engineering, University Of Arizona .
- [14] Nelson, M, The Data Compression Book, 2nd Edition Publication, 1996 .
- [15] William K. Pratt, "Digital Image Processing", John Wiley And Sons, Inc., Second Edition, New York, 1991.
- [16] Raghuveer M. Rao And Ajit S. Bapardikar, "Wavelet Transforms", Addison - Wesley, First Edition, 2000.
- [17] H.R. Mahadevaswamy, P. Janardhanan and Y. Venkataramani, "Lossless Image Compression using Wavelets - A Comparative Study", Proceedings of National Communication Conference, NCC'99, IIT Khargpur, India, pp.345-352, Allied Publishers, NewDelhi, January 1999.
- [18] A.R. Calderbank, Ingrid Dauchies, Wim Sweldens, and Boon lock Yeo, "Lossless Image Compression using Integer to Integer Wavelet

Transforms”, International conference on Image Processing, ICIP’97, Vol 1, No.385, pp.596-599, Oct 1997.

- [19] Amir Said, and William A. Pearman, “An Image Multiresolution Representation For Lossless And Lossy Compression”, IEEE Transactions on Image Processing, Vol.5, pp.1303-1310, Sept.1996.
- [20] Marc Antonini, Miche, Barland, Pierre Mathieu, and Ingrid Daubechies, “Image Coding Using Wavelet Transform”, IEEE Transactions on Image Processing, Vol.1, No:2, pp.205-220 ,April 1992.
- [21] Osma K.Al-Shaykh, and Russell M.Mersereau, “Lossy Compression of Noisy Images”,IEEE Transactions on Image Processing, Vol.7, No.12, pp.1641-1652, December 1998.
- [22] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi “The JPEG 2000 Still ImageCompression Standard “ in “IEEE Signal processing”, Sep 2001.
- [23] Arun N. Netravali and Barry G. Haskell, “Digital Pictures-Representation, Compression, and Standards”, Plenum Press, New York, 2nd Edition, 1994.
- [24] S. Golomb, “Run Length Encodings”, IEEE Transactions on Information Theory, Volume IT-12, pp.399-401, 1986.
- [25] Peggy Morton and Karin Glinden , “Image Compression using the haar wavelet transform”, Plenum Press, October 1996.
- [26] Peggy Morton & Arne Petersen , “Image Compression using the haar wavelet transform “, Math 45 – College of the Redwoods , December 1997.
- [27] LiHong Huang Herman , “Image Compression using the haar wavelet transform, Plenum Press, December 13, 2001.
- [28] <http://www.spelman.edu/~colm/>
- [29] http://online.redwoods.cc.ca.us/instruct/darnold/fall_2002/ames/
- [30] Proc. IEEE(special issue on wavelets), vol 84, Apr. 1996.
- [31] N. Jayant & P. Noll, Digital Coding waveforms: Principles & Applications to Speech & Video Englewood Cliffs, NJ. Prentice Hall, 1984.
- [32] B. Zovko – Cihlar, S. Grgic & D. Modric , “ Coding techniques in multimedia communication “ I Proc 2nd Int. Workshop Image & Signal Process IWISP ’95 , Budapest, Hungary, 1995.



Fig. 3: Original and Haar Wavelet Transformed images for Different Levels



Fig. 4: Reconstructed Images for different Levels

Information Security and Ethics in Educational Context: Propose a Conceptual Framework to Examine Their Impact

Hamed Taherdoost

Islamic Azad University, Semnan
Branch
Department of Computer
Tehran, Iran
hamed.taherdoost@gmail.com

Meysam Namayandeh

Islamic Azad University, Islamshahr
Branch
Department of Computer
Tehran, Iran
meysam.namayandeh@gmail.com

Neda Jalaliyoon

Islamic Azad University, Semnan
Branch
Department of Management
Semnan, Iran
neda.jalaliyoon@yahoo.com

Abstract— Information security and ethics are viewed as major areas of interest by many academic researchers and industrial experts. They are defined as an all-encompassing term that refers to all activities needed to secure information and systems that supports it in order to facilitate its ethical use. In this research, the important parts of current studies introduced. To accomplish the goals of information security and ethics, suggested framework discussed from educational level to training phase in order to evaluate computer ethics and its social impacts. Using survey research, insight is provided regarding the extent to which and how university student have dealt with issues of computer ethics and to address the result of designed computer ethics framework on their future career and behavioral experience.

Keywords—component; information security; ethics; framework

I. INTRODUCTION

The current development in information and communication technologies impacted all sectors in our daily life. To ensure effective working of information security factors, various controls and measures had been implemented by current policies and guidelines between computer developers [7]. However, lack of proper computer ethics studies in this field motivated researcher s to define a new framework.

Hence, this research will examine awareness and information of students in computer ethics from educational aspect. Also from Malaysian perspective, review of related research [11] indicates the existence of conflicting views concerning the ethical perceptions of students. In today's global economy, computer security and computer ethics awareness is an important component of any management information system [13].

It be would an undeniable element of security in Malaysian computer technology as Malaysia is ranked 8 out of 10 top infected countries in the Asia Pacific region as a target for cyber attackers [14]. Indeed, points out that there is a need to understand the basic cultural, social, legal and ethical issues inherent in the discipline of computing. For such reasons, it would be important that future computer professionals are taught the meaning of responsible conduct [9].

As the computer ethics was one of the major topics which have been throughout the past decades, in this part of introduction we reviewed a short milestone on computer ethics and related history of developments. During the late 1970s, Joseph Weizenbaum, a computer scientist at Massachusetts Institute of Technology in Boston, created a computer program that he called ELIZA. In his first experiment with ELIZA, he scripted it to provide a crude imitation of a psychotherapist engaged in an initial interview with a patient. In the mid 1970s, Walter Maner began to use the term "computer ethics" to refer to that field of inquiry dealing with ethical problems aggravated, transformed or created by computer technology.

Maner offered an experimental course on the subject at University. During the late 1970s, Maner generated much interest in university-level computer ethics courses. He offered a variety of workshops and lectures at computer science conferences and philosophy conferences across America.

By the 1980s, a number of social and ethical consequences of information technology were becoming public issues in the world, issues like computer-enabled crime, disasters caused by computer failures, invasions of privacy via computer databases, and major law suits regarding software ownership. Because of the work of Parker and others, the foundation had been laid for computer ethics as an academic discipline. In the mid-80s, James Moor of Dartmouth College published his influential

article "What is Computer Ethics? In Computers and Ethics, a special issue of the journal on that particular time.

During the 1990s, new university courses, research centers, conferences, journals, articles and textbooks appeared, and a wide diversity of additional scholars and topics became involved. The mid-1990s has heralded the beginning of a second generation of Computer Ethics which contain the new concept of security. The time has come to build upon and elaborate the conceptual foundation whilst, in parallel, developing the frameworks within which practical action can occur, thus reducing the probability of unforeseen effects of information technology application.

In 2000s, the computer revolution can be usefully divided into three stages, two of which have already occurred, the introduction stage and the permeation stage.

The world entered the third and most important stage "the power stage" in which many of the most serious social, political, legal, and ethical questions involving information technology will present them on a large scale. The important mission in this era is to believe that future developments in information technology will make computer ethics more vibrant and more important than ever. Computer ethics is made to research about security and its beneficial aspects.

The remainder of this paper is organized as follows: section 2 describes the details of DAMA framework by further phases on section 3. In section 4 the related theories are discussed from ethical views.

II. FRAMEWORK

This research is going to propose a framework for development of information security with computer ethics respect to educational conception. The further discussion follows the exact code of ethics which are including Privacy, Property, Accuracy and Accessibility. As Figure 1 depicts, DAMA (Delimma, Attitude, Morality, and Awareness) framework examines information security and computer ethics from two major dimensions: the educational and security training. In addition, DAMA framework are also explored to suggested the educational core of computer ethics which is the effective ways to teach information security along with computer ethics from the basis of educational level rather than higher level.

The educational dimension is focusing on the core of information security which considers along with awareness, morality, attitude and dilemma. In fact, educational dimension is explored from various perspectives to have relevance for group rather than individuals where the main focus of this issue has been mentioned in training level. Examples of questions in order to guide the development of DAMA framework references include: have you ever heard about computer ethics? What are ethical dilemmas and its social impacts?

The other main phase of educational dimension is moral development that includes personal beliefs related to their background of computer ethics. In fact, it focus on morality and further effectiveness that how individual morality can change

their attitude and therefore acquire appropriate awareness hence evaluate ethical dilemmas.

Moreover, security and training dimension is what students themselves manifest core of information security along with the help of formal and informal discussion. The security dimension includes informal discussion of common mistakes that happens among most of security consultant and officers which are relevant to information security ethics. It includes discussions of specific exploits of current weaknesses and may result as unethical behavior. The goal of security dimension is to communicate students from technical perspective to theoretical training.

DAMA approaches present methods and creative ideas for teaching of computer ethics with respect of information security for diverse audiences. The framework's dimensions cover the basic levels for computer ethics lectures and class room discussions related to ethical behavior of future computer scientists. The main emphasis is to presents creative and beneficial methods for learning experiences in various kinds of information security ethics. The authors place particular focus that will require students to build and rebuilt their beliefs in different ways in order to know unethical behaviors and their social impact on their future career.

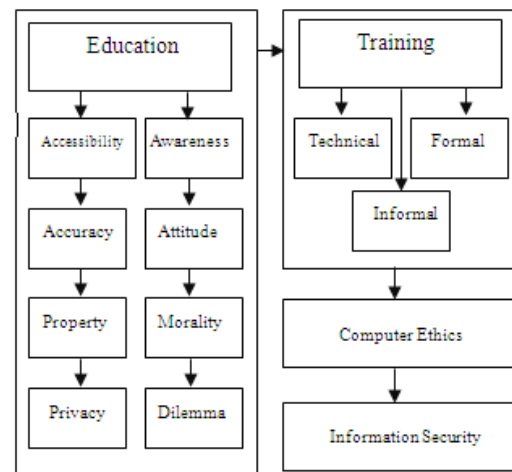


Figure 1. DAMA Framework

III. EDUCATIONAL DIMENSION

A. DAMA

Computer education now begins in elementary school and is no longer a restricted technical specialty learned only by those who are going to design or program computers. Because of the widespread prevalence of computers in society a core of ethical precepts relating to computer technology should be communicated not only to computer professionals, but to the general public through all levels of education. The issue should be viewed from the perspective of society and perspective of computer professionals [15].

In looking at the computer ethics there is a great emphasis upon incorporating ethical and social impact issues throughout

the curriculum starting at the point when children first become computer users in school. In particular, there are a set of guidelines regarding what students in general need to know about computer ethics. The preparation of future computer professionals should be examined at both the high school and university computer science curriculum [4]. The researchers [11] are in the process of developing new recommendations at both levels of curriculum. In the high school curriculum, there will be both general and specific approaches to ethics and social impact issues.

The general approach is to incorporate these concerns across the curriculum, not just in computer courses. This is in keeping with the philosophy that computers should be integrated across the curriculum as a tool for all disciplines. The specific approach is to develop social impact modules within the computer courses that will focus on these concerns ([5], 2004). At the university level the researchers faces a yet-to-be resolved dilemma of how to implement the proposed societal strand in the new curriculum recommendations. There is much discussion, but little action, regarding the necessity of preparing ethically and socially responsible computer scientists, especially in light of the highly publicized computer viruses that are an embarrassment to the profession.

When combined with other computer science core material, the teaching of ethics is made complicated by the fact that it is not as concrete as the rest of the curriculum. In accepting the value-laden nature of technology, researchers should recognize the need to teach a methodology of explicit ethical analysis in all decision-making related technology. The moral development is at the heart of interest in the morality element. In this model [3], researchers wanted to create educational opportunities that allow students to examine their existing beliefs regarding ethical and technical issues and in relation to existing technical, professional, legal, and cultural solutions. In an earlier section, it described how students examine these solutions with an external, objective point of view.

Now, the student is positioned at the centre of the intersecting circles. The aim is to create educational opportunities that allow and encourage students to explore "who am I now" in relation to technical, professional, cultural, and legal solutions to these ethical and security issues, and asks questions such as "what is the relationship between who am I, who I want to be, and these issues and solutions"? The most important factor in effective computer security is people's attitudes, actions, and their sense of right and wrong [8]. Problems and issues raised in the computing environment, Topics to be discussed include misuse of computers, concepts of privacy, codes of conduct for computer professionals, disputed rights to products, defining ethical, moral, and legal parameters, and what security practitioners should do about ethics.

The issue of computer security has fallen into the gray area that educators and industry alike have avoided for fear that too little knowledge could be hazardous and too much could be dangerous. Most organizations acknowledge the need for data security, but, at the same time, approach security as hardware. It may be more important, and far more successful to address

the issue of data security as an attitude rather than a technology.

B. PAPA

According to [12] decision makers place such a high value on information that they will often invade someone's privacy to get it. Marketing researchers have been known to go through people's garbage to learn what products they buy, and government officials have stationed monitors in restrooms to gather traffic statistics to be used in justifying expansion of the facilities.

These are examples of snooping that do not use the computer. The general public is aware that the computer can be used for this purpose, but it is probably not aware of the ease with which personal data can be accessed. If you know how to go about the search process, you can obtain practically any types of personal and financial information about private citizens. Here four major aspect of Mason's theory shall be studied:

1) Privacy

Privacy may define as the claim of individuals to determine for themselves when, to whom, and to what extent individually identified data about them is communicated or used. Most invasions of privacy are not this dramatic or this visible. Rather, they creep up on us slowly as, for example, when a group of diverse files relating to a student and his or her activities are integrated into a single large database. Collections of information reveal intimate details about a student and can thereby deprive the person of the opportunity to form certain professional and personal relationships.

This is the ultimate cost of an invasion of privacy. So why integrate databases in the first place. It is because the bringing together of disparate data makes the development of new information relationships possible.

2) Accuracy

Accuracy represents the legitimacy, precision and authenticity with which information is rendered. Because of the pervasiveness of information about individuals and organizations contained in information systems, special care must be taken to guard against errors and to correct known mistakes. Difficult questions remain when inaccurate information is shared between computer systems. Any framework should describe the legal liability issues associated with information. Who is held accountable for the errors? This is an important question may come across every researcher's mind or which party liable for inexact or incorrect information that leads to devastation of another.

3) Property

One of the more controversial areas of computer ethics concerns the intellectual property rights connected with software ownership. Some people, like Richard Stallman who started the Free Software Foundation, believe that software ownership should not be allowed at all. He claims that all information should be free, and all programs should be available for copying, studying and modifying by anyone who wishes to do so. Others argue that software companies or programmers would not invest weeks and months of work and

significant funds in the development of software if they could not get the investment back in the form of license fees or sales [12].

Today's software industry is a multibillion dollar part of the economy; and software companies claim to lose billions of dollars per year through illegal copying. Many people think that software should be own able, but "casual copying" of personally owned programs for one's friends should also be permitted. The software industry claims that millions of dollars in sales are lost because of such copying.

4) *Accessability*

Accessability represents the legitimacy, precision and authenticity with which information is rendered. Regarding this important aspect of research this question may come across the people's mind who is held accountable for errors? Who can you trust in order to outsource your project? In fact, in term computer ethics accessability means, what kind of information would be available for the legal users and students.

IV. SECURITY AND TRAINING LEVEL

In terms of computer ethics, security would be an undeniable factor of it. Therefore, short review on information security which is influence in computer ethics will help the researcher to identify the further study. Many different terms have been used to describe security in the IT areas where information security has become a commonly used concept, and is a broader term than data security and IT security. Information is dependent on data as a carrier and on IT as a tool to manage the information. Information security is focused on information that data represent, and on related protection requirements.

So the definition of information system security is "the protection of information systems against unauthorized access to or modification of information, whether in storage, processing or transit, and against the denial of service to authorized users or the provision of service to unauthorized users, including those measures necessary to detect, document, and counter such threats". Four characteristics of information security are: availability, confidentiality, integrity and accountability, simplified as "the right information to the right people in the right time". *Availability*: concerns the expected use of resources within the desired timeframe. *Confidentiality*: relates to data not being accessible or revealed to unauthorized people. *Integrity*: concerns protection against undesired changes. *Accountability*: refers to the ability of distinctly deriving performed operations from an individual. Both technical and administrative security measures are required to achieve these four characteristics.

A. *TECHNICAL LEVEL SECURITY*

From a technical perspective, the preservation of confidentiality, integrity availability and accountability requires the adoption of IT security solutions such as encryption of data and communication, physical eavesdropping, access control systems, secure code programming, authorization and authentication mechanisms, database security mechanisms, intrusion detection systems, firewalls. At this level it is possible

to introduce frameworks and methods for the selection of the appropriate technological solution depending on the needs for a particular application with respect to security in computer ethics.

B. *Formal level security*

The formal level of information security is related with the set of policies, rules, controls, standards, etc. aimed to define an interface between the technological subsystem (Technical level) and the behavioral (computer ethics) subsystem (Informal level).

According to many definitions of an information security, this is the level where much of the effort of the information security is concentrated. An interesting review of the security literature identifies a trend in information system research moving away from a narrow technical viewpoint towards a socio-organizational perspective.

C. *Informal level security*

In the domain of the informal level of information security, the unit of analysis is individual and the research is concerned about behavioral issues like values, attitude, beliefs, and norms that are dominant, and influencing an individual employee regarding security practices in an organization. The solutions suggested in this domain are more descriptive than prescriptive in nature and the findings at this level need to be effectively implemented through other levels (i.e. formal and technical). An interesting review of research papers in the behavioral or computer ethical domain is, looking at used theories, suggested solutions, current challenges, and future research [1].

V. THEORIES PERSPECTIVE

Ethics is an important facet of comprehensive security of information system's security. Research in ethics and information systems has been also carried outside the information security community. Anyhow, researcher sees that the relationship of hackers and information security personnel has not yet been properly analyzed. Within this short review, a philosophical point of view shall be taken, and problems of establishing ethical protection measures against violations of information security shall be studied. Further analysis leads to quite opposite results of the main stream arguments that support the need of common ethical theories for information security. This addition provides with a framework that is feasible within the current technology, supports natural social behavior of human beings and is iterative enabling forming of larger communities from smaller units.

Recently, the trend appears to be that the ethics approved by the security community is having the law enforcement [2]. Several attempts around the world are made to enforce proper behavior in the information society by theoretical methods. From information security point of view, hackers are seen as criminals, unaware of the results of their immoral activities making fun out of serious problems.

Hacker community, on the other hand, sees information security staff as militants that respecting the freedom of

individual and information [6]. Further depth into the conflict can be found by introducing another dimension to the classification of ethical theories into two categories: Phenomenologist vs. Positivist and individualist vs. collectivist ethics.

Phenomenologism vs. Positivism: According to the phenomenological school, what is good is given in the situation, derived from the logic and language of the situation or from dialogue and debate about “goodness”. Positivism encourages s to observe the real world and derive ethical principles inductively.

Individualism vs. Collectivism: According to the individualistic school, the moral authority is located in the individual whereas collectivism says that a larger collectivity must care the moral authority. Major schools, based on these concepts, can be listed to be Collective Rule-Based Ethics, Individual Rule- Based Ethics. A detailed analysis of these schools is provided by [10].

Also from distributed information systems perspective security of information systems requires both technical and non-technical measures, special effort must be paid on the assurance that all methods support each other and do not set contradictory or infeasible requirements for each other which contain two major theoretical elements:

Ethics negotiation phase is where organizations or individuals representing themselves negotiate the content of ethical communication agreement over specific communication channels.

Ethics enforcement phase is where each organization enforces changes in the ethical code of conduct by specifying administrative and managerial routines, operational guide lines, monitoring procedures and sanctions for unacceptable behavior. Organizations or university individuals involved in negotiation should code desired ethical norms in terms of acceptable behavior within the information processing. Agreement should be searched and once reached, contract made and agreed norms enforced throughout the organization. In the optimal case, ethics has the law enforcement and juridical actions against violations can be prosecuted in court.

VI. CONCLUSION

Educational centers within higher educational level have unique opportunity to help and educate computer users in order to face with ethical dilemmas. Therefore, this would be the main challenge of this study to focus on computer ethics with the help of suggested framework. As a result, computer ethics

is becoming a field in need of research based upon a necessity to provide information for education which is related to security concepts. The legal structure appears to be limited in its ability to provide ethical behavior effectively. While not wishing to be alarmists, research suggests the needs to be concerted effort on the part of the all the computer professional societies to update their ethical codes and to incorporate a process of continual security.

REFERENCES

- [1] Bynum, T., Computer ethics: Basic concepts and historical overview, Stanford, Encyclopedia of Philosophy. 2006.
- [2] Cruz, J., and Frey, W., An effective strategy for integrating ethics across the curriculum in engineering, An ABET 2000 Challenge, Science and Engineering Ethics, vol. 9, no. 3, pp. 543-568, 2004.
- [3] Dark, M., Epstein, R., Morales, L., Counterline, T., Yuan, Q., Ali, M., Rose, M., and Harter, N., A framework for information security ethics education, Proc. Of the 10th Colloquium for Information Systems Security Education, University of Maryland, University College Adelphi, MD June 5-8, 2006.
- [4] Forcht, K. A., Pierson, J. K., and Bauman, B. M., Developing awareness of computer ethics, ACM, 1998.
- [5] Foster, A. L., Insecure and unaware, The Chronicle of Higher Education, (May 7, 2004), p. 33.
- [6] Fowler, T. B., Technology's changing role in intellectual property rights, IT Pro4, vol.2, pp. 39-44, 2004.
- [7] Hamid, N., Information security and computer ethics: Tools, theories and modeling, North Carolina University , Igbi Science Publication, vol. 1, pp. 543-568, 2007
- [8] Huff, C., and Frey, W., Good computing: A pedagogically focused model of virtue in the practice of computing, Under Review, pp. 30-32, 2005.
- [9] Langford, D., Practical computer ethics, London: McGraw Hill, pp. 118-127, 2000.
- [10] Leiwo, J., and Heikkuri, S., An analysis of ethics as foundation of information security in distributed systems, Proc. 31st Annual Hawaii International Conf. on System Sciences, pp. 213-222, 1998.
- [11] Maslin, M., and Zuraini, I., Computer security and computer ethics awareness: A component of management information system, Malaysia Conf. IEEE Technology and Society Magazine, 2008.
- [12] Mason, R. O., Four ethical issues of the information age, Management Information Systems Quarterly, vol. 10, no. 1, pp. 5-12, 1986.
- [13] North, M. M., George, R., and North, S. M. Computer security and ethics awareness in university environments: A challenge for management of information systems, ACM, Florida, United States of America, pp. 434-439, 2006.
- [14] Sani, R., Cybercrime Gains Momentum, New Straits Times, April 3, 2006
- [15] Spinello, R., Cyberethics: morality and law in cyberspace, Third edition, Sudbury, vol. 2, 2003.

Finding Fuzzy Locally Frequent Itemsets

Fokrul Alom Mazarbhuiya
Department of Computer Science
College of Computer Science
King Khalid University
Abha, Kingdom of Saudi Arabia
e-mail: fokrul_2005@yahoo.com

Md. Ekramul Hamid
Department of Network Engineering
College of Computer Science
King Khalid University
Abha, Kingdom of Saudi Arabia
e-mail: ekram_hamid@yahoo.com

Abstract— The problem of mining temporal association rules from temporal dataset is to find association between items that hold within certain time intervals but not throughout the dataset. This involves finding frequent sets that are frequent at certain time intervals and then association rules among the items present in the frequent sets. In fuzzy temporal datasets as the time of transaction is imprecise, we may find set of items that are frequent in certain fuzzy time intervals. We call these as fuzzy locally frequent sets and the corresponding associated association rules as fuzzy local association rules. These association rules cannot be discovered in the usual way because of fuzziness involved in temporal features. In this paper, we propose a modification to the well-known A-priori algorithm to compute fuzzy locally frequent sets. Finally we have shown manually with the help of an example that the algorithm works.

Keywords- Temporal Data mining, Fuzzy number, Fuzzy time-stamp, Core length of a fuzzy interval, Fuzzified interval

I. INTRODUCTION

The problem of mining association rules has been defined initially [15] by R. Agarwal *et al* for application in large super markets. Large supermarkets have large collection of records of daily sales. Analyzing the buying patterns of the buyers will help in taking typical business decisions such as what to put on sale, how to put the materials on the shelves, how to plan for future purchase etc.

Mining for association rules between items in temporal databases has been described as an important data-mining problem. Transaction data are normally temporal. The market basket transaction is an example of this type.

In this paper we consider datasets, which are fuzzy temporal i.e. the time in which a transaction has taken place is imprecise or approximate and is attached to the transactions. In large volumes of such data, some hidden information or relation ship among the items may be there which cannot be extracted because of some fuzziness in the temporal features. Also the case may be that some association rules may hold in certain fuzzy time period but not throughout the dataset. For finding such association rules we need to find itemsets that are frequent at certain time period, which will obviously be imprecise due to the fact that the time of each transaction is fuzzy. We call such frequent sets fuzzy locally frequent over fuzzy time interval. From these fuzzy locally frequent sets, associations among the items can be obtained. It is shown manually with the help of an example that the algorithm gives the required result. Although it is assumed here that the fuzzy

time stamps are having similar membership functions, we claim that the same algorithm with slight modification can be applied to the database having dissimilar fuzzy time stamps.

In section II we give a brief discussion on the works related to our work. In section III we describe the definitions, terms and notations used in this paper. In section IV, we give the algorithm proposed in this paper for mining fuzzy locally frequent sets. In section V, we explain the algorithm with a small dataset and display the results. We conclude with conclusion and lines for future work in section VI. In the last section we give some references.

II. RELATED WORKS

The problem of discovery of association rules was first formulated by Agrawal *et al* in 1993. Given a set I , of items and a large collection D of transactions involving the items, the problem is to find relationships among the items i.e. the presence of various items in the transactions. A transaction t is said to support an item if that item is present in t . A transaction t is said to support an itemset if t supports each of the items present in the itemset. An association rule is an expression of the form $X \Rightarrow Y$ where X and Y are subsets of the itemset I . The rule holds with confidence τ if $\tau\%$ of the transaction in D that supports X also supports Y . The rule has support σ if $\sigma\%$ of the transactions supports $X \cup Y$. A method for the discovery of association rules was given in [15], which is known as the A priori algorithm. This was then followed by subsequent refinements, generalizations, extensions and improvements. As the number of association rules generated is too large, attempts were made to extract the useful rules ([13], [16]) from the large set of discovered association rules. Attempts are also made to make the process of discovery of rules faster ([12], [14]). Generalized association rules ([9], [17]) and Quantitative association rules ([18]) were later on defined and algorithms were developed for the discovery of these rules. A hashed based technique is used in [11] to improve the rule mining process of the A priori algorithm.

Temporal Data Mining is now an important extension of conventional data mining and has recently been able to attract more people to work in this area. By taking into account the time aspect, more interesting patterns that are time dependent can be extracted. There are mainly two broad directions of temporal data mining [7]. One concerns the discovery of causal relationships among temporally

oriented events. Ordered events form sequences and the cause of an event always occur before it. The other concerns the discovery of similar patterns within the same time sequence or among different time sequences. The underlying problem is to find frequent sequential patterns in the temporal databases. The name sequence mining is normally used for the underlying problem. In [8] the problem of recognizing frequent episodes in an event sequence is discussed where an episode is defined as a collection of events that occur during time intervals of a specific size.

The association rule discovery process is also extended to incorporate temporal aspects. In temporal association rules each rule has associated with it a time interval in which the rule holds. The problems associated are to find valid time periods during which association rules hold, the discovery of possible periodicities that association rules have and the discovery of association rules with temporal features. In [10], [19], [20] and [21], the problem of temporal data mining is addressed and techniques and algorithms have been developed for this. In [10] an algorithm for the discovery of temporal association rules is described. In [2], two algorithms are proposed for the discovery of temporal rules that display regular cyclic variations where the time interval is specified by user to divide the data into disjoint segments like months, weeks, days etc. Similar works were done in [6] and [22] incorporating multiple granularities of time intervals (e.g. first working day of every month) from which both cyclic and user defined calendar patterns can be achieved. In [1], the method of finding locally and periodically frequent sets and periodic association rules are discussed which is an improvement of other methods in the sense that it dynamically extract all the rules along with the intervals where the rules hold. In ([23], [24]) fuzzy calendric data mining and fuzzy temporal data mining is discussed where user specified ill-defined fuzzy temporal and calendric patterns are extracted from temporal data.

Our approach is different from the above approaches. We are considering the fact that the time of transactions are not precise rather they are fuzzy numbers and some items are seasonal or appear frequently in the transactions for certain ill-defined periods only i.e. summer, winter, etc. They appear in the transactions for a short time and then disappear for a long time. After this they may again reappear for a certain period and this process may repeat. For these itemsets the support cannot be calculated in the usual way ([1], [10]), it has to be computed by the method defined in section 3B. These items may lead to interesting association rules over fuzzy time intervals. In this paper we calculate the support values of these sets locally in a α -cut of a fuzzy time interval where a fuzzy time interval represents a particular season in which the itemset is appearing frequently and if they are frequent in the fuzzy time interval under consideration then we call these sets fuzzy locally frequent sets. The large fuzzy time gap in which they do not appear is not counted. As mentioned in

the previous paragraph similarly works were also done in [23], [24] but in non-fuzzy temporal data. But in all these methods they discuss the association rule mining of non-fuzzy temporal data. Our approach although little bit similar to the work of [1], is different from others in the sense that it discovers association rules from fuzzy temporal data and finds the association rules along with their fuzzy time intervals over which the rules hold automatically.

III. PROBLEM DEFINITION

A. Some Definition related to Fuzziness

Let E be the universe of discourse. A fuzzy set A in E is characterized by a membership function $A(x)$ lying in $[0,1]$. $A(x)$ for $x \in E$ represents the grade of membership of x in A . Thus a fuzzy set A is defined as

$$A = \{(x, A(x)), x \in E\}$$

A fuzzy set A is said to be normal if $A(x) = 1$ for at least one $x \in E$

An α -cut of a fuzzy set is an ordinary set of elements with membership grade greater than or equal to a threshold α , $0 \leq \alpha \leq 1$. Thus a α -cut A_α of a fuzzy set A is characterized by

$$A_\alpha = \{x \in E; A(x) \geq \alpha\} \text{ [see e.g. [4]]}$$

A fuzzy set is said to be convex if all its α -cuts are convex sets [see e.g. [5]].

A fuzzy number is a convex normalized fuzzy set A defined on the real line R such that

1. there exists an $x_0 \in R$ such that $A(x_0) = 1$, and
2. $A(x)$ is piecewise continuous.

A fuzzy number is denoted by $[a, b, c]$ with $a < b < c$ where $A(a) = A(c) = 0$ and $A(b) = 1$. $A(x)$ for all $x \in [a, b]$ is known as left reference function and $A(x)$ for $x \in [b, c]$ is known as the right reference function. Thus a fuzzy number can be thought of as containing the real numbers within some interval to varying degrees. The α -cut of the fuzzy number $[t_1 - a, t_1, t_1 + a]$ is a closed interval $[t_1 + (\alpha - 1) \cdot a, t_1 + (1 - \alpha) \cdot a]$.

Fuzzy intervals are special fuzzy numbers satisfying the following.

1. there exists an interval $[a, b] \subset R$ such that $A(x_0) = 1$ for all $x_0 \in [a, b]$, and
2. $A(x)$ is piecewise continuous.

A fuzzy interval can be thought of as a fuzzy number with a flat region. A fuzzy interval A is denoted by $A = [a, b, c, d]$ with $a < b < c < d$ where $A(a) = A(d) = 0$ and $A(x) = 1$ for all $x \in [b, c]$. $A(x)$ for all $x \in [a, b]$ is known as left reference function and $A(x)$ for $x \in [c, d]$ is known as the right reference function. The left reference function is non-decreasing and the right reference function is non-increasing [see e.g. [3]].

Similarly the α -cut of the fuzzy interval $[t_1-a, t_1, t_2, t_2+a]$ is a closed interval $[t_1+(\alpha-1).a, t_2+(1-\alpha).a]$.

The core of a fuzzy number A is the set of elements of A having membership value one i.e.

$$\text{Core}(A) = \{x, A(x); A(x) = 1\}$$

For every fuzzy set A ,

$$A = \bigcup_{\alpha \in [0,1]} \alpha A$$

where $\alpha A(x) = \alpha \cdot A(x)$, and αA is a special fuzzy set [4]

For any two fuzzy sets A and B and for all $\alpha \in [0, 1]$,

$$\text{i) } \alpha(A \cup B) = \alpha A \cup \alpha B$$

$$\text{ii) } \alpha(A \cap B) = \alpha A \cap \alpha B$$

For any two fuzzy numbers A and B , we say the membership functions $A(x)$ and $B(x)$ are similar to each other if the slope of the left reference function of $A(x)$ is equal to the that of $B(x)$ and the slope of right reference of $A(x)$ is equal that of $B(x)$. Obviously for any two fuzzy numbers A and B having similar membership functions

$$|\alpha A| = |\alpha B|, \forall \alpha \in [0, 1]$$

B. Some Definition related to Fuzzy Locally Frequent set

Let $T = \langle t_0, t_1, \dots \rangle$ be a sequence of imprecise or fuzzy time stamps over which a linear ordering $<$ is defined where $t_i < t_j$ means t_i denotes the core of a fuzzy time which is earlier than the core of another fuzzy time stamp t_j . For the sake of convenience, we assume that all the fuzzy time stamps are having similar membership functions. Let I denote a finite set of items and the transaction database D is a collection of transactions where each transaction has a part which is a subset of the itemset I and the other part is a fuzzy time-stamp indicating the approximate time in which the transaction had taken place. We assume that D is ordered in the ascending order of the core of fuzzy time stamps. For fuzzy time intervals we always consider a fuzzy closed intervals of the form $[t_1-a, t_1, t_2, t_2+a]$ for some real number a . We say that a transaction is in the fuzzy time interval $[t_1-a, t_1, t_2, t_2+a]$ if the α -cut of the fuzzy time stamp of the transaction is contained in α -cut of $[t_1-a, t_1, t_2, t_2+a]$ for some user's specified value of α .

We define the local support of an itemset in a fuzzy time interval $[t_1-a, t_1, t_2, t_2+a]$ as the ratio of the number of transactions in the time interval $[t_1+(\alpha-1).a, t_2+(1-\alpha).a]$ containing the itemset to the total number of transactions in $[t_1+(\alpha-1).a, t_2+(1-\alpha).a]$ for the whole data base D for a given

value of α . We use the notation $\text{Sup}_{[t_1-a, t_1, t_2, t_2+a]}(X)$ to denote the support of the itemset X in the fuzzy time interval $[t_1-a, t_1, t_2, t_2+a]$. Given a threshold σ we say that an itemset X is frequent in the fuzzy time interval $[t_1-a, t_1, t_2, t_2+a]$ if

$\text{Sup}_{[t_1-a, t_1, t_2, t_2+a]}(X) \geq (\sigma/100) \cdot \text{tc}$ where tc denotes the total number of transactions in D that are in the fuzzy time interval

$[t_1-a, t_1, t_2, t_2+a]$. We say that an association rule $X \Rightarrow Y$, where X and Y are item sets holds in the time interval $[t_1-a, t_1, t_2, t_2+a]$ if and only if given threshold τ ,

$$\text{Sup}_{[t_1-a, t_1, t_2, t_2+a]}(X \cup Y) / \text{Sup}_{[t_1-a, t_1, t_2, t_2+a]}(X) \geq \tau / 100.0$$

and $X \cup Y$ is frequent in $[t_1-a, t_1, t_2, t_2+a]$. In this case we say that the confidence of the rule is τ .

For each locally frequent item set we keep a list of fuzzy time intervals in which the set is frequent where each fuzzy interval is represented as $[start-a, start, end, end+a]$ where $start$ gives the approximate starting time of the time interval and end gives the approximate ending time of the time-interval. $end - start$ gives the length of the core of the fuzzy time interval. For a given value of α of two intervals $[start_1-a, start_1, end_1, end_1+a]$ and $[start_2-a, start_2, end_2, end_2+a]$ are non-overlapping if their α -cuts are non-overlapping.

IV. PROPOSED ALGORITHM

A. Generating Fuzzy Locally Frequent Sets

While constructing locally frequent sets, with each locally frequent set a list of fuzzy time-intervals is maintained in which the set is frequent. Two user's specified thresholds α and $minthd$ are used for this. During the execution of the algorithm while making a pass through the database, if for a particular itemset the α -cut of its current fuzzy time-stamp, $[^{\alpha}L_{current}, ^{\alpha}R_{current}]$ and the α -cut, $[^{\alpha}L_{lastseen}, ^{\alpha}R_{lastseen}]$ of its fuzzy time, when it was last seen overlap then the current transaction is included in the current time-interval under consideration which is extended with replacement of $^{\alpha}R_{lastseen}$ by $^{\alpha}R_{current}$; otherwise a new time-interval is started with $^{\alpha}L_{current}$ as the starting point. The support count of the item set in the previous time interval is checked to see whether it is frequent in that interval or not and if it is so then it is fuzzified and added to the list maintained for that set. Also for the fuzzy locally frequent sets over fuzzy time intervals, a minimum core length of the fuzzy period is given by the user as $minthd$ and fuzzy time intervals of core length greater than or equal to this value are only kept. If $minthd$ is not used than an item appearing once in the whole database will also become locally frequent a over fuzzy point of time.

Procedure to compute L_1 , the set of all fuzzy locally frequent item sets of size 1.

For each item while going through the database we always keeps an α -cut $^{\alpha}lastseen$ which is $[^{\alpha}L_{lastseen}, ^{\alpha}R_{lastseen}]$ that corresponds to the fuzzy time stamp when the item was last seen. When an item is found in a transaction and the fuzzy time-stamp is tm and if its α -cut $^{\alpha}tm = [^{\alpha}L_{tm}, ^{\alpha}R_{tm}]$ has empty intersection with $[^{\alpha}L_{lastseen}, ^{\alpha}R_{lastseen}]$, then a new time interval is started by setting $start$ of the new time interval as $^{\alpha}L_{tm}$ and end of the previous time interval as $^{\alpha}R_{lastseen}$. The previous time interval is fuzzified provided the support of the item is greater than $min-sup$. The fuzzified interval is then added to the list maintained for that item provided that the duration of the core is greater than $minthd$. Otherwise $^{\alpha}R_{lastseen}$ is set to $^{\alpha}R_{tm}$, the counters maintained for counting

transactions are increased appropriately and the process is continued.

Following is the algorithm to compute L_1 , the list of locally frequent sets of size 1. Suppose the number of items in the dataset under consideration is n and we assume an ordering among the items.

- Algorithm1

$C_1 = \{(i_k, tp[k]) : k = 1, 2, \dots, n\}$

where i_k is the k -th item and $tp[k]$ points to a list of fuzzy time intervals initially empty.

for $k = 1$ to n do

set ${}^{\alpha}lastseen[k] = \phi$;

set $itemcount[k]$ and $transcount[k]$ to zero for each transaction t in the database with fuzzy time stamp tm

do

{for $k = 1$ to n do

{ if $\{i_k\} \subseteq t$ then

{ if ${}^{\alpha}lastseen[k] = \phi$

{ ${}^{\alpha}lastseen[k] = {}^{\alpha}firstseen[k] = {}^{\alpha}tm$;

$itemcount[k] = transcount[k] = 1$;

}

else

if $({}^{\alpha}Llastseen[k], {}^{\alpha}Rlastseen[k]) \cap$

$[{}^{\alpha}Ltm[k], {}^{\alpha}Rtm[k]] \neq \phi$

{ ${}^{\alpha}Rlastseen[k] = {}^{\alpha}Rtm[k]$; $itemcount[k]++$;

$transcount[k]++$;

}

else

{ if $(itemcount[k]/transcount[k]*100 \geq \sigma)$

fuzzify $([{}^{\alpha}Llastseen[k], {}^{\alpha}Rlastseen[k]], \forall \alpha$

$\in [0, 1])$

if $(|core(fuzzified interval)| \geq minthd)$

add $(fuzzified interval)$ to $tp[k]$;

$itemcount[k] = transcount[k] = 1$;

$lastseen[k] = firstseen[k] = tm$;

}

}

else $transcount[k]++$;

} // end of k -loop //

} // end of do loop //

for $k = 1$ to n do

{ if $(itemcount[k]/transcount[k]*100 \geq \sigma)$

fuzzify $([{}^{\alpha}Llastseen[k], {}^{\alpha}Rlastseen[k]], \forall \alpha \in [0, 1])$

if $(|core(fuzzified interval)| \geq minthd)$

add $(fuzzified interval)$ to $tp[k]$;

if $(tp[k] \neq 0)$ add $\{i_k, tp[k]\}$ to L_1

}

fuzzify $([{}^{\alpha}a, {}^{\alpha}b], \alpha)$

$$\bigcup_{\alpha \in [0,1]} [{}^{\alpha}a, {}^{\alpha}b]$$

where ${}_{\alpha}[a, b](x) = \alpha \cdot {}^{\alpha}[a, b](x)$

return $(fuzzified interval)$

}

Two support counts are kept, $itemcount$ and $transcount$. If the count percentage of an item in an α -cut of a fuzzy time interval is greater than the minimum threshold then only the set is considered as a locally frequent set over fuzzy time interval.

L_1 as computed above will contain all 1-sized locally frequent sets over fuzzy time intervals and with each set there is associated an ordered list of fuzzy time intervals in which the set is frequent. Then A priori candidate generation algorithm is used to find candidate frequent set of size 2. With each candidate frequent set of size two we associate a list of fuzzy time intervals that are obtained in the pruning phase. In the generation phase this list is empty. If all subsets of a candidate set are found in the previous level then this set is constructed. The process is that when the first subset appearing in the previous level is found then that list is taken as the list of fuzzy time intervals associated with the set. When subsequent subsets are found then the list is reconstructed by taking all possible pair wise intersection of subsets one from each list. Sets for which this list is empty are further pruned.

Using this concept we describe below the modified A-priori algorithm for the problem under consideration.

- Algorithm2

Modified A priori

Initialize

$k = 1$;

C_1 = all item sets of size 1

L_1 = {frequent item sets of size 1 where

with each itemset $\{i_k\}$ a list $tp[k]$ is maintained which gives all time fuzzy

intervals in which the set is frequent}

L_1 is computed using algorithm 1.1 */

for $(k = 2; L_{k-1} \neq \phi; k++)$ do

{ C_k = apriorigen(L_{k-1})

/ same as the candidate generation method of the A priori algorithm setting $tp[i]$ to zero for all i */*

prune(C_k);

drop all lists of fuzzy time intervals maintained with the sets in C_k

Compute L_k from C_k .

// L_k can be computed from C_k using the same procedure used for computing L_1 //

$k = k + 1$

}

Answer = $\bigcup_k L_k$

Prune(C_k)

{Let m be the number of sets in C_k and let the sets be $s_1, s_2, \dots, s_m$. Initialize the pointers $tp[i]$ pointing to the list of fuzzy time-intervals maintained with each set s_i to null

for $i = 1$ to m do

{for each $(k-1)$ subset d of s_i do

{if $d \notin L_{k-1}$ then

$\{C_k = C_k - \{s_i, tp[i]\}; break;\}$

else

{ if $(tp[i] == \text{null})$ then set $tp[i]$ to point to the list of fuzzy time intervals maintained for d

else

{ take all possible pair-wise intersection of fuzzy time intervals one from each list, one list maintained with $tp[i]$ and the other maintained with d and take this as the list for $tp[i]$

delete all fuzzy time intervals whose core length is less than the value of $minthd$ if $tp[i]$ is empty then $\{C_k = C_k - \{s_i, tp[i]\};$

break;

}

}

}

}

}

}

V. EXPLANATION OF THE ALGORITHM WITH EXAMPLE

To illustrate the above algorithms, we consider a dataset of two days consisting of fuzzy time stamps and set of transactions. Here each fuzzy time stamp is associated with a transactions means that the transaction occurs at a fuzzy time. For the sake of convenience, we take all the fuzzy time stamps as triangular fuzzy numbers. The dataset is given below:

Fuzzy time-stamps	set of transactions
[0, 2, 4]	1 3 4 5
[1, 3, 5]	1 4 6 7 9
[2, 4, 6]	2 3 5 6 7 10
[3, 5, 7]	1 3 5 7 10
[4, 6, 8]	2 3 4 8 9
[5, 7, 9]	1 2 4 5 6 10
[6, 8, 10]	1 2 3 4 7 8 10
[7, 9, 11]	1 2 3 4 5 6 7 8
[8, 10, 12]	1 2 3 4 5 7 8 9
[9, 11, 13]	2 3 4 6 7 10
[10, 12, 14]	1 2 3 4 5 7 10
[11, 13, 15]	1 2 8 10
[12, 14, 16]	1 2 3 4 5 7
[13, 15, 17]	2 3 6 7
[14, 16, 18]	1 2 3 4 5
[15, 17, 19]	1 2 3 4
[16, 18, 20]	2 5 6 7
[17, 19, 21]	1 2 3 4 5
[18, 20, 22]	2 4 8 10
[19, 21, 23]	2 3 6 9
[20, 22, 24]	3 4 7 8
[21, 23, 1']	1 6 10
[22, 24, 2']	1 5
[23, 1', 3']	7
[24, 2', 4']	2 10
[1', 3', 5']	3 4
[2', 4', 6']	1 4 5 8 9 10
[3', 5', 7']	1 2 4
[4', 6', 8']	2 4 5
[5', 7', 9']	2 3 4 6 7
[6', 8', 10']	3 5
[7', 9', 11']	1 3 4 5 6 7
[8', 10', 12']	2 5
[9', 11', 13']	1 2 3 7 8
[10', 12', 14']	1 9 10
[11', 13', 15']	1 2 3 6 10
[12', 14', 16']	2 3 4
[13', 15', 17']	1 2 3
[14', 16', 18']	1 2 3
[15', 17', 19']	3 4 5 7 8
[16', 18', 20']	1 2 3 8 10
[17', 19', 21']	2 3 5
[18', 20', 22']	1 2 3 4 5 6
[19', 21', 23']	1 2 3 5 7 8
[20', 22', 24']	2 3 4 5 9 10

We execute the algorithm manually with the above dataset taking $min-sup = 0.4$ and $minthd = 3$.

After first pass we have the set of 1-item frequent sets along with the fuzzy intervals where they are frequent as

$L_1 = \{(\{1\}; [0, 2, 19, 21], [7', 9', 21', 23']),$
 $(\{2\}; [2, 4, 21, 23], [8', 10', 22', 24']),$
 $(\{3\}; [0, 2, 22, 24], [5', 7', 22', 24']),$
 $(\{4\}; [4, 6, 22, 24], [1', 3', 9', 11']),$
 $(\{5\}; [0, 2, 19, 21], [2', 4', 10', 12']),$
 $[15', 17', 22', 24']),$

({6}; [5, 7, 11, 13]),
({7}; [6, 8, 15, 17], [5', 7', 11', 13']),
({8}; [4, 6, 10, 12]),
({10}; [2, 4, 8, 10])

Candidates for the second pass are $C_2 = \{\{1\ 2\}, \{1\ 3\}, \{1\ 4\}, \{1\ 5\}, \{1\ 6\}, \{1\ 7\}, \{1\ 8\}, \{1\ 10\}, \{2\ 3\}, \{2\ 4\}, \{2\ 5\}, \{2\ 6\}, \{2\ 7\}, \{2\ 8\}, \{2\ 10\}, \{3\ 4\}, \{3\ 5\}, \{3\ 6\}, \{3\ 7\}, \{3\ 8\}, \{3\ 10\}, \{4\ 5\}, \{4\ 6\}, \{4\ 7\}, \{4\ 8\}, \{4\ 10\}, \{5\ 6\}, \{5\ 7\}, \{5\ 8\}, \{5\ 10\}, \{6\ 7\}, \{6\ 8\}\}$

After the second pass, we got the second level frequent sets as

$L_2 = \{\{1\ 2\}; [5, 7, 19, 21], [9', 11', 21', 23']\},$
({1 3}; [6, 8, 19, 21], [7', 9', 21', 23']),
({1 4}; [5, 7, 19, 21]),
({1 5}; [3, 5, 16, 18]),
({1 7}; [6, 8, 14, 16]),
({1 10}; [3, 5, 8, 10]),
({2 3}; [2, 4, 21, 23], [9', 11', 22', 24']),
({2 4}; [4, 6, 20, 22]),
({2 5}; [7, 9, 19, 21], [17', 19', 22', 24']),
({2 6}; [5, 7, 11, 13]),
({2 7}; [6, 8, 15, 17]),
({2 8}; [4, 6, 10, 12]),
({3 4}; [4, 6, 19, 21]),
({3 5}; [0, 2, 5, 7], [7, 9, 16, 18],
[15', 17', 22', 24']),
({3 7}; [6, 8, 15, 17], [5', 7', 11', 13']),
({3 8}; [4, 6, 10, 12]),
({4 5}; [5, 7, 16, 18]),
({4 7}; [6, 8, 14, 16]),
({4 8}; [4, 6, 10, 12]),
({5 7}; [7, 9, 14, 16]),
({5 10}; [2, 4, 7, 9])

Candidates for the third pass are

$C_3 = \{\{1\ 2\ 3\}, \{1\ 2\ 4\}, \{1\ 2\ 5\}, \{1\ 2\ 7\}, \{1\ 3\ 4\}, \{1\ 3\ 5\}, \{1\ 3\ 7\}, \{1\ 4\ 7\}, \{1\ 5\ 7\}, \{2\ 3\ 4\}, \{2\ 3\ 5\}, \{2\ 3\ 7\}, \{2\ 3\ 8\}, \{2\ 4\ 5\}, \{2\ 4\ 7\}, \{2\ 4\ 8\}, \{2\ 5\ 7\}, \{3\ 4\ 5\}, \{3\ 4\ 7\}, \{3\ 4\ 8\}, \{3\ 5\ 7\}, \{4\ 5\ 7\}, \{4\ 5\ 8\}\}$

After the third pass, we got the third level frequent sets as

$L_3 = \{\{1\ 2\ 3\}; [6, 8, 19, 21], [9', 11', 21', 23']\},$
({1 2 4}; [5, 7, 19, 21]),
({1 2 5}; [5, 7, 16, 18]),
({1 2 7}; [6, 8, 14, 16]),
({1 3 4}; [6, 8, 19, 21]),
({1 3 5}; [7, 9, 16, 18]),
({1 3 7}; [6, 8, 14, 16]),
(1 4 7); [6, 8, 14, 16]),
({1 5 7}; [7, 9, 14, 16]),
({2 3 4}; [4, 6, 19, 21]),
({2 3 5}; [7, 9, 16, 18]),
({2 3 7}; [6, 8, 15, 17]),
({2 3 8}; [4, 6, 10, 12]),
({2 4 5}; [5, 7, 16, 18]),
({2 4 7}; [6, 8, 14, 16]),
({2 4 8}; [4, 6, 10, 12]),
({2 5 7}; [7, 9, 14, 16]),
({3 4 5}; [7, 9, 16, 18]),
({3 4 7}; [6, 8, 14, 16]),

({3 4 8}; [4, 6, 10, 12]),
({3 5 7}; [7, 9, 14, 16]),
({4 5 7}; [7, 9, 14, 16])

Candidates for the fourth pass are

$C_4 = \{\{1\ 2\ 3\ 4\}, \{1\ 2\ 3\ 5\}, \{1\ 2\ 3\ 7\}, \{1\ 2\ 4\ 5\}, \{1\ 2\ 4\ 7\}, \{1\ 2\ 5\ 7\}, \{1\ 3\ 4\ 5\}, \{1\ 3\ 4\ 7\}, \{1\ 3\ 5\ 7\}, \{2\ 3\ 4\ 5\}, \{2\ 3\ 4\ 7\}, \{2\ 3\ 4\ 8\}, \{2\ 3\ 5\ 7\}, \{2\ 4\ 5\ 7\}, \{2\ 4\ 5\ 7\}\}$

$L_4 = \{\{1\ 2\ 3\ 4\}; [6, 8, 19, 21]\},$

({1 2 3 5}; [6, 8, 16, 18]),
({1 2 3 7}; [6, 8, 14, 16]),
({1 2 4 5}; [5, 7, 16, 18]),
({1 2 4 7}; [6, 8, 14, 16]),
({1 2 5 7}; [6, 8, 14, 16]),
({1 3 4 5}; [7, 9, 16, 18]),
({1 3 4 7}; [6, 8, 14, 16]),
({1 3 5 7}; [7, 9, 14, 16]),
({2 3 4 5}; [7, 9, 16, 18]),
({2 3 4 7}; [6, 8, 15, 17]),
({2 3 4 8}; [4, 6, 10, 12]),
({2 3 5 7}; [7, 9, 15, 17]),
({2 4 5 7}; [6, 8, 14, 16]),
({3 4 5 7}; [7, 9, 12, 14])

Candidates for the fifth pass are

$C_5 = \{\{1\ 2\ 3\ 4\ 5\}, \{1\ 2\ 3\ 4\ 7\}, \{1\ 2\ 3\ 5\ 7\}, \{1\ 2\ 4\ 5\ 7\}, \{1\ 3\ 4\ 5\ 7\}, \{2\ 3\ 4\ 5\ 7\}\}$

After the fifth pass, we got fifth level frequent sets as

$L_5 = \{\{1\ 2\ 3\ 4\ 5\}; [7, 9, 16, 18]\},$
({1 2 3 4 7}; [6, 8, 14, 16]),
({1 2 3 5 7}; [7, 9, 14, 16]),
({1 2 4 5 7}; [7, 9, 14, 16]),
({1 3 4 5 7}; [7, 9, 14, 16]),
({2 3 4 5 7}; [7, 9, 14, 16])

Candidates for sixth pass are $C_6 = \{\{1\ 2\ 3\ 4\ 5\ 7\}\}$

After the sixth pass we got sixth level frequent sets as

$L_6 = \{\{1\ 2\ 3\ 4\ 5\ 7\}; [7, 9, 14, 16]\}$
Answer = $\{\{1\ 2\ 3\ 4\ 5\ 7\}; [7, 9, 14, 16]\},$
({2 3 4 8}; [4, 6, 10, 12]),
({1 2}; [9', 11', 21', 23']),
({1 3}; [7', 9', 21', 23']),
({1 10}; [3, 5, 8, 10]),
({2 3}; [9', 11', 22', 24']),
({2 5}; [17', 19', 22', 24']),
({3 5}; [15', 17', 22', 24']),
({3 7}; [5', 7', 11', 13']),
({5 10}; [2, 4, 7, 9]),
({4}; [1', 3', 9', 11']),
({5}; [2', 4', 10', 12'])

B. Generating Association Rules

If an itemset is frequent in a fuzzy time-interval $[t_1 - a, t_1, t_2, t_2 + a]$ then all its subsets are also frequent in the fuzzy time-interval $[t_1 - a, t_1, t_2, t_2 + a]$. But to generate the association rules as defined in section 3, we need the supports of the subsets in fuzzy time-interval $[t_1 - a, t_1, t_2, t_2 + a]$, which may not be available after application of the algorithm as defined in 4.1. For this one more scan of the whole database will be needed.

For each association rule we attach a fuzzy interval in which the association rule holds.

CONCLUSIONS AND LINES FOR FUTURE WORKS

An algorithm for finding frequent sets that are frequent in certain fuzzy time periods from fuzzy temporal data, is given in the paper. The algorithm dynamically computes the frequent sets along with their fuzzy time intervals where the sets are frequent. These frequent sets are named as fuzzy locally frequent sets. The technique used is similar to the A priori algorithm. From these fuzzy locally frequent sets interesting rules may follow.

In the level-wise generation of fuzzy locally frequent sets, for each fuzzy locally frequent set we keep a list of all fuzzy time-intervals in which it is frequent. For generating candidates for the next level, pair-wise intersections of the intervals in two lists are taken. Further, we tested manually with an example that the algorithm works. For the sake convenience, we have taken here the dataset having fuzzy time stamps with similar membership functions, but the algorithm can be applicable to the dataset with dissimilar fuzzy time stamps. The same algorithm can be implemented with both real life as well as synthetic datasets.

REFERENCES

- [1] A. K. Mahanta, F. A. Mazarbhuiya and H. K. Baruah; Finding Locally and Periodically Frequent Sets and Periodic Association Rules, Proceeding of 1st Int'l Conf on Pattern Recognition and Machine Intelligence (PreMI'05), LNCS 3776, 576-582, 2005.
- [2] B. Ozden, S. Ramaswamy and A. Silberschatz; Cyclic Association Rules, Proc. of the 14th Int'l Conference on Data Engineering, USA, 412-421, 1998.
- [3] D. Dubois and H. Prade; Ranking fuzzy numbers in the setting of possibility theory, *Information Science* 30, 183-224, 1983.
- [4] G. J. Klir and B. Yuan; Fuzzy Sets and Fuzzy Logic Theory and Applications, Prentice Hall India Pvt. Ltd., 2002.
- [5] G. Q. Chen, S. C. Lee and E. S. H. Yu; Application of fuzzy set theory to economics, In: Wang, P. P., ed., *Advances in Fuzzy Sets, Possibility Theory and Applications*, Plenum Press, N. Y., 277-305, 1983.
- [6] G. Zimbrão, J. Moreira de Souza, V. Teixeira de Almeida and W. Araújo da Silva; An Algorithm to Discover Calendar-based Temporal Association Rules with Item's Lifespan Restriction, Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (2002) Canada, 2nd Workshop on Temporal Data Mining, v. 8, 701-706, 2002.
- [7] J. F. Roddick, M. Spillopoulou; A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research; ACM SIGKDD, June 1999.
- [8] H. Manilla, H. Toivonen and I. Verkamo; Discovering frequent episodes in sequences; KDD'95; AAAI, 210-215, August 1995.
- [9] J. Hipp, A. Myka, R. Wirth and U. Guntzer; A new algorithm for faster mining of generalized association rules; Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98), Nantes, France, (September 1998).
- [10] J. M. Ale and G.H. Rossi; An approach to discovering temporal association rules; Proceedings of the 2000 ACM symposium on Applied Computing, March 2000.
- [11] J. S. Park, M. S. Chen and P. S. Yu; An Effective Hashed Based Algorithm for Mining Association Rules; Proceedings of ACM SIGMOD, 175-186, 1995.
- [12] M. J. Zaki, S. Parthasarathy, M. Ogihara and W. Li; New algorithms for the fast discovery of association rules; Proceedings of the 3rd

International Conference on KDD and data mining (KDD '97), Newport Beach, California, August 1997.

- [13] M. Klemettinen, H. Manilla, P. Ronkainen, H. Toivonen and A. I. Verkamo; Finding interesting rules from large sets of discovered association rules; Proceedings of the 3RD international Conference on Information and Knowledge Management, Gathersburg, Maryland, 29 Nov 1994.
- [14] R. Agrawal and R. Srikant; Fast algorithms for mining association rules, Proceedings of the 20th International Conference on Very Large Databases (VLDB '94), Santiago, Chile, June 1994.
- [15] R. Agrawal, T. Imielinski and A. Swami; Mining association rules between sets of items in large databases; Proceedings of the ACM SIGMOD '93, Washington, USA, May 1993.
- [16] R. Motwani, E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, J. D. Ullman and C. Yang; Finding interesting association rules without support pruning, Proceedings of the 16th International Conference on Data Engineering (ICDE), IEEE, 2000.
- [17] R. Srikant and R. Agrawal; Mining generalized association rules, Proceedings of the 21st Conference on very large databases (VLDB '95), Zurich, Switzerland, September 1995.
- [18] R. Srikant and R. Agrawal; Mining quantitative association rules in large relational tables, Proceedings of the 1996 ACM SIGMOD Conference on management of data, Montreal, Canada, June 1996.
- [19] X. Chen and I. Petrounias; A framework for Temporal Data Mining; Proceedings of the 9th International Conference on Databases and Expert Systems Applications, DEXA '98, Vienna, Austria. Springer-Verlag, Berlin; Lecture Notes in Computer Science 1460, 796-805, 1998.
- [20] X. Chen and I. Petrounias; Language support for Temporal Data Mining; Proceedings of 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD '98, Springer Verlag, Berlin, 282-290, 1998.
- [21] X. Chen, I. Petrounias and H. Healthfield; Discovering temporal Association rules in temporal databases; Proceedings of IADT'98 (International Workshop on Issues and Applications of Database Technology, 312-319, 1998.
- [22] Y. Li, P. Ning, X. S. Wang and S. Jajodia; Discovering Calendar-based Temporal Association Rules, In Proc. of the 8th Int'l Symposium on Temporal Representation and Reasoning, 2001.
- [23] W. J. Lee and S. J. Lee; Discovery of Fuzzy Temporal Association Rules, IEEE Transactions on Systems, Man and Cybernetics-part B; Cybernetics, Vol 34, No. 6, 2330-2341, Dec 2004.
- [24] W. J. Lee and S. J. Lee; Fuzzy Calendar Algebra and Its Applications to Data Mining, Proceedings of the 11th International Symposium on Temporal Representation and Reasoning (TIME'04), IEEE, 2004.

AUTHOR'S PROFILE



Fokrul Alom Mazarbhuiya received B.Sc. degree in Mathematics from Assam University, India and M.Sc. degree in Mathematics from Aligarh Muslim University, India. After this he obtained the Ph.D. degree in Computer Science from Gauhati University, India. Since 2008 he has been serving as an Assistant Professor in College of Computer Science, King Khalid University, Abha, kingdom of Saudi Arabia. His research interest includes Data Mining, Information security, Fuzzy Mathematics and Fuzzy logic



Md. Ekramul Hamid received his B.Sc and M.Sc degree from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh. After that he obtained the Masters of Computer Science degree from Pune University, India. He received his PhD degree from Shizuoka University, Japan. During 1997-2000, he was a lecturer in the Department of Computer Science and Technology, Rajshahi University. Since 2007, he has been serving as an associate professor in the same department. He is currently working as an assistant professor in the college of computer science at King Khalid University, Abha, KSA. His research interests include speech enhancement, and speech signal processing.

An Extensible Cloud Architecture Model for Heterogeneous Sensor Services

R.S.Ponmagal

Dept of CSE
Anna University of Technology
Tiruchirappalli, India
rsponmagal@gmail.com

J.Raja

Dept of ECE
Anna University of Technology
Tiruchirappalli, India
raja@tau.edu.in

Abstract— This paper aims to examine how the sensor information can be shared, through a new resource called “cloud”. The recent research issue in integrating Wireless sensor Networks with Cloud is to establish a faster communication link between the two. The wireless sensor networks are used to sense and collect information required. The sensor information is deployed into the cloud through the sensor profile for web services. The Sensor Profile for Web Services specifies a lightweight subset of the overall Web services protocol suite that is appropriate for network-connected sensors. Cloud generally offers resources on demand. Since wireless sensor networks are limited in their processing power, battery life and communication speed, cloud computing usually offers the opposite, which makes it attractive for long term observations, analysis and use in different kinds of environments. In this paper a model is presented, which combines the concept of wireless sensor networks with the cloud computing paradigm, and show how both can benefit from this combination. Sensor data access is thus moved from loosely managed system to a well managed cloud. The integration of sensor information into the cloud through the sensor as a service paradigm proves the faster communication establishment. The scalability of this approach seems to be unlimited, since wireless sensor networks operate independently, and are connected to the cloud computing environment through a scalable number of wireless sensor network communication gateways. The cloud computing environment itself offers a scalable infrastructure, which makes it very attractive. The main goal is to design a flexible architecture in which sensor network’s information can be accessed by applications efficiently.

Keywords- cloud, sensor profiles, webservices, sensors, resources, information, scalability.

I. INTRODUCTION

Wireless Sensor Networks consists of energy constrained sensor nodes and a Sink node with higher processing capabilities. The sensors are physically composed of electronic sensing circuitry, a processor and a wireless transceiver, plus a power supply unit (battery). Sensor networks are distributed event based systems that focus on simple data gathering applications and operate notably differently from that of traditional computer networks. The gathered data can be made accessible to other nodes, including a specialized one called sink through a variety of means. TOSSIM is used to generate

sensor data. The future sensor networks are envisioned as comprising heterogeneous devices assisting to a large range of applications. Interoperability is required for such heterogeneous devices. To achieve this, we propose a Service Oriented approach for the data acquisition from sensor network, and an extensible architecture in which this web services based deployment is extended to CLOUD. Here the sensor nodes are service providers and applications are clients of such services. Hosting a web service challenges battery life, bandwidth, processing power constraints of low power sensor nodes.

The language in which the sensors are proposed to speak is Sensor profiles for web services. The sensor information is also planned to be deployed into the cloud through the sensor profile for web services. The Sensor Profile for Web Services specifies a lightweight subset of the overall Web services protocol suite that is appropriate for network-connected sensors. The proposed Sensor profiles for web services reduce the data processing at sensor nodes, while keeping the complex data processing at sink. The sensor profiles reduces the power consumption of sensor nodes, hence maximizes the network life time. The Sensor Profile prescribes how to use elements of core Web services specifications to enable these functions: Send more secure messages to and from a Web service, Dynamically discover a Web service, Describe a Web service Subscribe to, and receive events from, a Web service. Complete set of functionalities for sensor integration and limited constrained functionalities of sensors can be specified. It further reduces the interdependencies between the sensors. The huge amount of data, which a sensor network is able to deliver, demands a powerful and scalable storage and processing infrastructure. Depending on the sample frequency (e.g. from 100 Hz or more down to few samples a day for calculating observations) of the sensors, the deployed infrastructure has to scale up memory, storage and processing power. Today wireless sensor network platforms (e.g. TOSSIM, Crossbow MicaZ, Sentilla JCreate, SunSpot) that perform sensing and complex calculations are most of the time constrained in their capabilities and therefore is one appropriate way to solve this issue is to do offline processing of sensor data if the resources are not sufficient.

In this paper a model is presented, which combines the concept of wireless sensor networks with the cloud

computing paradigm, and show how both can benefit from this combination. Sensor data access is thus moved from loosely managed system to a well managed cloud. The integration of sensor information into the cloud through the sensor as a service through sensor profiles for web services languages will prove the faster communication establishment.

The scalability of this approach seems to be unlimited, since wireless sensor networks operate independently, and are connected to the cloud computing environment through a scalable number of wireless sensor network communication gateways. The cloud computing environment itself offers a scalable infrastructure, which makes it very attractive. Hence, the sensed information is deployed into the STAX, cloud architecture. The combination of wireless sensor networks, with their huge amount of gathered sensor data and their limited processing power, with a cloud computing infrastructure makes it attractive in terms of *i)* integration of sensor network platforms from different vendors, *ii)* scalability of data storage, *iii)* scalability of processing power for different kinds of analysis, *iv)* worldwide access to the processing and storage infrastructure, *v)* resource optimization, *vi)* be able to share the results more easily, and *vii)* using pricing as one more criteria for the IT infrastructure.

The present work defines the proposed architectural components as well as the protocol stack of the SPWS middleware. The paper is organized as follows: Section II covers the state of the art. Section III describes the related work. Section IV,V,VI and VII details the system architecture and the related components. Section VIII shows the performance analysis. Section IX outlines the conclusion.

II. STATE OF THE ART

The sensor information can be transmitted to the requesting client as [1] SOAP messages, which is used to access the sensed information with application independent protocol. A service approach for the design of wireless sensor networks is explained. Services are defined as the data provided by sensor nodes and the applications to be executed on those data. Clients access the sensor network by submitting queries to those services.

The DPWS proposal is optimized as TinyDPWS [2] with application specific format technique, reduces the energy consumed by the sensor nodes. An advanced middleware solution to the problem of integrating a Wireless Sensor Network into the information system of an enterprise at a high abstraction level. This is achieved by using the proposed middleware which provides to the wireless sensors a Service Oriented Architecture connection to the Internet. The proposed middleware is based on the Device Profile for Web Services which is a Service Oriented Architecture technology at the device level.

A method to access the sensor information using structured data [3] and WSDL descriptions is proposed. The functionality and data provided by the new nodes is exposed in a structured manner, so that multiple applications may access them. The result is a highly inter-operable system where multiple applications can share a common evolving sensor

substrate. A key challenge in using web services on resource constrained sensor nodes is the energy and bandwidth overhead of the structured data formats used in web services.

Integrating wireless sensor networks in heterogeneous networks is a complex task. A reason is the absence of a standardized data exchange format that is supported in all participating sub networks. XML has evolved to the de facto standard data exchange format between heterogeneous networks and systems. However, XML usage within sensor networks has not been introduced because of the limited hardware resources. In this paper, an XML template objects are introduced making XML usage applicable within sensor networks. Different optimized way [4] of using XML is specified. This new XML data binding technique provides significant high compression results while still allowing dynamic XML processing and XML navigation.

The standard device profiles for web services [5] which could be used for wireless sensor networks is proposed. Even if DPWS is the best candidate to integrate WSN in existing infrastructures, it cannot be applied to WSN without research efforts, because it addresses softer resource constraints as required in WSN. But DPWS provides a minimal set of constraints for applications in resource constrained devices. So this paper describes an approach that further restricts DPWS for WSNs, but keeps it still interoperable with DPWS.

A cloud storage platform [6] for pervasive computing environments such as wireless sensor networks is explained. Data storage and sharing is difficult for these sensors due to the data inflation and the natural limitations, such as the limited storage space and the limited computing capability. Since the emerging cloud storage solutions can provide reliable and unlimited storage, they satisfy to the requirement of pervasive computing very well. Thus a new cloud storage platform is designed which includes a series of shadow storage services to address these new data management challenges in pervasive computing environments, which called as "SmartBox".

An efficient way of combining cloud computing and wireless sensor networks [7] is explained. The cloud provides scalable processing power and several kinds of connectable services. This distributed architecture has many similarities with a typical wireless sensor network, where a lot of nodes, which are responsible for sensing and local preprocessing, are interconnected with wireless connections. Since wireless sensor networks are limited in their processing power, battery life and communication speed, cloud computing usually offers the opposite, which makes it attractive for long term observations, analysis and use in different kinds of environments.

Several service discovery protocols for wireless sensor networks [9] are proposed. In addition, to reduce power consumption we presented an activation schedule, based on the mapping of the nodes' operational modes to Bluetooth states. By announcing the activation schedule as a service, a representation of the state of the nodes is exposed to client applications. The proposed work takes into account of deploying the sensed data in STAX cloud using sensor profiles for web services approach.

III. SOA MODEL FOR SENSOR INFORMATION SYSTEM

Service Oriented Architectural model for representation of sensor services is shown in Fig.. The SOA model has three elements namely Sensor Service Provider, Sensor System Registry and Sensor Systems Client. The sensor system services are categorized in to Pressure sensing, Temperature sensing, and Level sensing services. A sensor Service provider offers the above services and describes the interface information of the services in interface description language called SensorSDL (Sensor Services Description Language) which is in the form of XML that makes the services available in the Sensor System Registry.

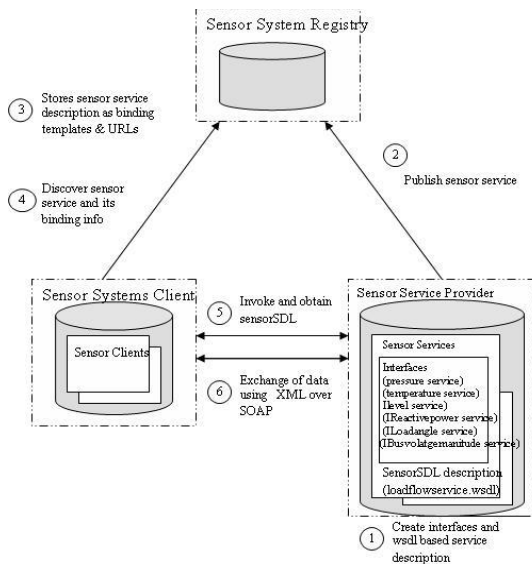


Figure 1.SOA model for sensor information system

Services are the key building blocks of SOA. A service is a reusable function that can be invoked by another component through a well-defined interface. Services are loosely coupled, that is, they hide their implementation details and only expose their interfaces. In this manner, sensor system client need not be aware of any underlying technology or programming language which the service is using. The loose coupling between services allows for a quicker response to changes than the existing conventional applications for sensor applications. This results in a much faster adoption to the need of applications which makes use of sensor applications.

The sensor system clients discover the service available in the registry by service names and acquire the interface information by SensorSDL of the sensor services. Based on this information, the clients have a binding with the sensor service provider and can invoke services using Simple Object Access Protocol (SOAP).

IV. THE EXTENSIBLE CLOUD ARCHITECTURE FOR SENSOR INFORMATION SYSTEM

A. Introduction

The Service Oriented Architectural model for sensor information is extended to cloud architecture. This paper proposes an advanced middleware solution to the problem of integrating a Wireless Sensor Network into the information system of a cloud at a high abstraction level. This is achieved by using the proposed middleware which provides to the wireless sensors a ServiceOriented Architecture connection to the Internet. The proposed middleware is based on the sensor Profile for Web Services which is a Service Oriented Architecture technology at the sensor level. Since this technology is based on exchanging eXtensible Markup Language documents, a technique is utilized which compresses and reduces the data volume of such documents at a level that can be handled by the use of the resource constrained environment of the wireless sensors. By utilizing the proposed middleware which implements only the basic functions of the sensor Profile for Web Services, we demonstrate how such a Wireless Sensor Network can be connected to the Cloud in which all its components conform to a Service Oriented Architecture standard.

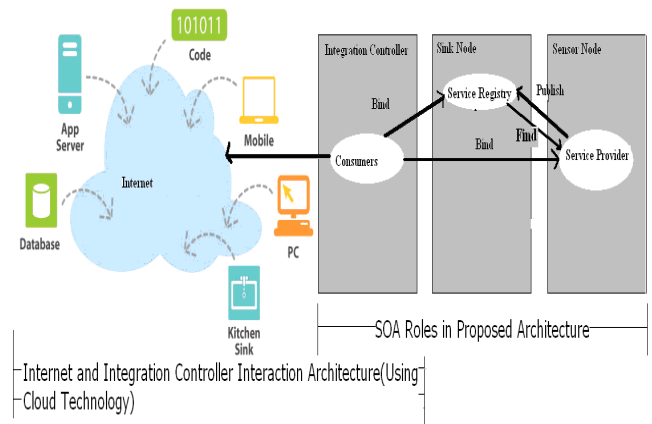


Figure 2. Proposed cloud architecture

B. System Design

1) *Sensor as software*

The sensor data is obtained from TOSSIM. TinyOS simulator is run on TinyOS1.7. NesC is the language used to simulate the sensor nodes. The TOSSIM itself got the packages to simulate real time sensors. Cygwin is used in windows platform to run the TOSSIM. In one cygwin window, the commands as in the Fig 3. is run. As a result a sensor node is simulated and its sensed parameters are written into tossim.txt file in the following path: C:\Program Files\UCB\cygwin\opt\tinyos-1.x\apps\Sense\tossim.txt. The sense folder also contains two NesC files called configuration (Sense.nc)file and Module(SenseM.nc)file.

In another cygwin window the appropriate commands are run, which opens up the tinyviz, which is builtin visualization tool available with TOSSIM.

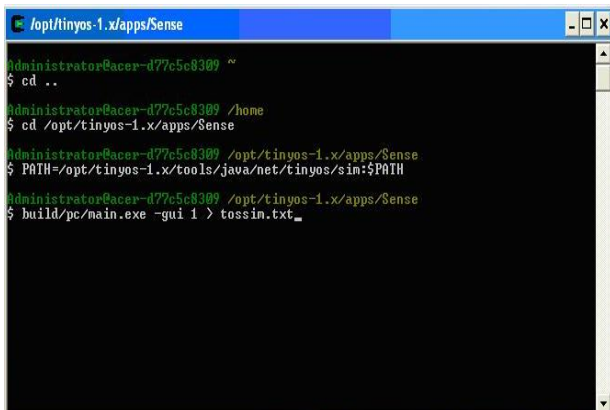


Figure 3. TOSSIM console

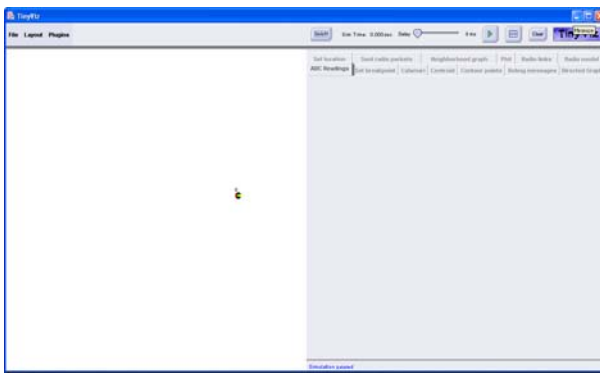


Figure 4. Simulation of a sensor node using TOSSIM

2) XML representation of Sensor data

Using XML as a standardized data exchange format in wireless sensor networks is a means to support more complex data management and heterogeneous networks. Moreover, XML is a key feature towards service-oriented sensor networks. Recent work has shown that XML can be compressed to meet the general hardware restrictions of sensor nodes while still supporting updates.

Integrating wireless sensor networks in heterogeneous networks is a complex task. A reason is that the absence of a standardized data exchange format that is supported in participating sub networks. Using xml in sensor networks encourages the interchangeability of different types of sensors and systems. The general verbosity of xml conflicts with the limited energy and memory capacities of sensor nodes. For this reason, native xml support has to be based on efficient data binding techniques that saves space, time and energy by eliminating the xml overhead. A good xml data binding solutions for sensor networks has to fulfill the following criteria:

Memory efficiency: representing high amount of xml data with a low amount of allocated memory. Runtime efficiency:

using only a minimum number of processing cycles for processing xml data. Processability: allowing to process xml data dynamically without an expensive decompressing step.

The following code listing specifies the temperature sensor's information:

```
<?xml version ="1.0"?>
<sensor>
<id>01</id>
<type>sensing</type>
<parameter>Temperature</parameter>
<units>56'kelvin</units>
<date>11</date>
<month>November</month>
<year>2009</year>
<time>11.00</time>
</sensor>
```

This is a step towards more complex but exchangeable data management in sensor networks and the extension of the service-oriented paradigm to sensor network application engineering.

Monitoring and controlling a physical environment has long been possible through device interfaces ranging from basic sensors and actuators to complex digital equipment and controllers. Such devices and the systems they enable have traditionally been the domain of embedded systems developers. We now see an increasing need and opportunity to create interfaces between the physical world of sensors and actuators and the software world of enterprise systems.

Wholesalers, retailers, and distributors demand immediate monitoring and control of shipments, enabled by RFID sensor data piped directly into their manufacturing, billing, and distribution software. Home healthcare monitoring can be implemented by providing devices such as ECG monitors and glucose monitors and pulse oximeters that can continuously monitor ambulatory patient status and alert healthcare providers of conditions requiring immediate care. A military control center must combine sensor data from various logistics and tactical environments—including the monitoring and control of RFID readers, vehicle control buses, GPS tracking systems, cargo climate controllers, and specialized devices—to provide situational awareness, preventive fleet maintenance, and real-time logistics.

The types of devices available for such scenarios continues to grow, while the cost of deploying them in the physical world and connecting them to all manner of networks continues to drop. However, the device interfaces, connections, and protocols are multiplying at a corresponding rate, and enterprise system developers are finding that integrating devices into the information technology (IT) world is daunting and expensive.

To eliminate much of the complexity and cost associated with integrating sensors into highly distributed enterprise systems, we propose leveraging existing and emerging standards from both the embedded-sensor and IT domains within a Service-Oriented Sensor Architecture (SOSA).

3) *Sensor profiles for web services*

At the moment, DPWS offers no trivial way to directly discover services available on the network. In a generic scenario, where a client does not know the services hosted on a device, a client always has to discover a device first and then discover its Hosted Services, with the help of the metadata provided in the description of a device. Bobek et al. [10] describe device and service templates for DPWS in a similar way to UPNP templates. The goal is to describe the service and

device types defined by DPWS at development time in a formal language and to shorten the discovery phase. The templates are divided into device and service templates, as DPWS differentiates between service and device types. Device templates describe the mandatory and optional services that must be implemented by devices offering a specific device type. Service templates describe a service that is related to a specific service type. Furthermore, device templates can include other device templates and build a hierarchical structure to enable extensibility. Bobek et al. relates the types transmitted during the discovery phase directly to a type defined by a device template. A service type that is part of the device description is related to a service template. The metadata, that is currently only available at runtime, can be formally defined at development time. With device and service templates, more static scenarios can be defined where a client already knows specific services and their endpoints when the client finds a device on the network. A typical example would be a printer device. This exemplary device type offers a printer service that is always hosted on port 80. The language proposed by Bobek et al. has some inconsistencies with the DPWS specification that are discussed and corrected in the next section. The template concept reduces the message exchanges on the network required to bind a client to a service also.

The device and service template concept proposed by Bobek et al. bases on a misunderstanding of the DPWS type model. WSDL port types are the recommended way to describe Web Service interfaces in WSDL documents. The service template concept relates DPWS service types to WSDL port types. As DPWS service types are directly related to WSDL port types, the service template concept is redundant. This direct relation is not clearly stated in the specification, but will be described in the non-normative documents that are published along with the DPWS 1.1 specification by OASISI WS-DD. Therefore, the template concept has to be revised in general. The template system is reduced to device templates only. A device template describes exactly one device type. Further device types may be included, to enable extensibility. The Hosted Service element may define a URL template for the endpoint reference of a Hosted Service. In contrast to the proposal by Bobek et al. this proposal refers directly to WSDL port types, as service types do in DPWS. To be in line with the

DPWS and WSDDiscovery specification, all types are lists of qualified names as specified in WS-Discovery.

4) *DPWS Enhancements for Wireless Sensor Networks*

DPWS allows the definition of application specific profiles. All enhancements, which are made in this section, are summarized in the new defined device type. Additional to the DPWS device type, a new device type for DPWS in WSN is introduced. The boxes in the sub sections are recommendations for requirements for the specification of the new WSN DPWS profile. The major goal of all restrictions and enhancements is the minimization of exchanged messages inside the WSN and the reduction of memory usage of DPWS implementations. The presented adaptations result in one discovery message only instead of the previously presented worst case scenario.

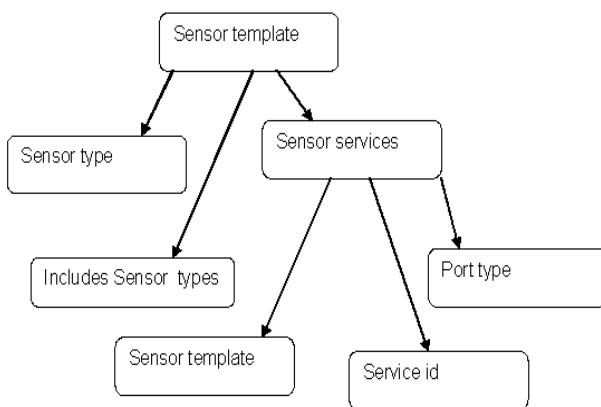
Templates for Devices Profile for Web Services: A language proposed by Bobek et al. to define service and device templates for DPWS. If sensor network nodes are modeled as DPWS devices (service providers), a sensor network is simply a heap of services. To implement more intelligent WSN, which require interaction inside WSN, sensor nodes must be modeled as peers that are DPWS devices and clients at the same time. Device templates can be used to create tailor-made DPWS clients for a specific scenario with smaller memory usage. At the same time, the template concept reduces the message exchanges on the network required to bind a client to a service also. The device and service template concept proposed by Bobek et al. bases on a misunderstanding of the DPWS type model. WSDL port types are the recommended way to describe Web Service interfaces in WSDL documents. The service template concept relates DPWS service types to WSDL port types. As DPWS service types are directly related to WSDL port types, the service template concept is redundant. This direct relation is not clearly stated in the specification, but will be described in the non-normative documents that are published along with the DPWS 1.1 specification by OASISI WS-DD. Therefore, the template concept has to be revised in general. The template system is reduced to device templates only. A device template describes exactly one device type. Further device types may be included, to enable extensibility. The Hosted Service element may define a URL template for the endpoint reference of a Hosted Service. In contrast to the proposal by Bobek et al. this proposal refers directly to WSDL port types, as service types do in DPWS. To be in line with the DPWS and WSDDiscovery [17] specification, all types are lists of qualified names as specified in WS-Discovery.

5) *conceptual and formal structure of sensor templates*

The following code listing contains an example where we define a sensor type for temperature detection.

```
<?xml version="1.0" ?>
<t:Relationship xmlns:t="http://www.ws4d
.org/templates/">
  <t:Host>
    <t:Type>
      <t:localName>sensor</t:localName>
    </t:Type>
```

```
<t:UrlTemplate>http://{ip}:4672/sensor</t:
UrlTemplate>
</t:Host>
<t:Hosted>
<t:Reference>
http://www.ws4d.org/templates/temperature/
temperatureService.wsdl
</t:Reference>
<t:ServiceId>
http://www.ws4d.org/temperature/
temperatureService
</t:ServiceId>
</t:Hosted>
<t:Hosted>
</t:Relationship>
```



SPWS sensor template		
Relationship		
1	Host	
	1	Type QName type
	*	Includes QName types
+	Hosted	
	XOR	1 Reference wsdl
		1 Hosted service
	?	Service id string

Figure 5. Sensor Templates

V. INTEROPERABILITY MODEL

An extension of the Service Oriented Architecture (SOA) paradigm to the device level is considered to comprise the technology that will provide the mapping of the resource constrained networks to the Internet. DPWS imposes high requirements on computing power and memory consumption so it is not designed to fit into resource constrained devices, such as wireless sensors. For this reason a special middleware for wireless sensors is proposed. This middleware implements

a stripped down version of the DPWS protocol stack which can fit into such devices. Furthermore a technique is proposed that allow the compression and reduction of the data volume of eXtensible Markup Language (XML) documents in such a way that they can be exchanged based on the Internet Protocol (IP) in the environment of the limited resources of WSN. Moreover, by enhancing the other components of the enterprise system with an extra mechanism that decompresses the compressed XML documents sent by sensors, we allow the communication between the sensors and the other components to be done in the complete SOA information exchange protocol at the

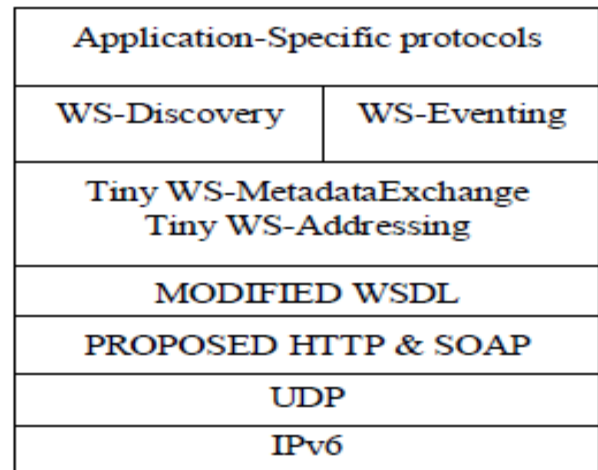


Figure 6. SPWS protocol stack

application level. The SPWS protocol stack, shown in Fig. 6, is proposed as a possible way of implementing SOA by the use Application-Specific protocols. It constitutes the proposed middleware and it is based on the traditional DPWS. Although it restricts the protocol stack of the traditional DPWS, it enables the wireless sensors to perform as they were hosting the traditional DPWS.

The first thing to consider when we develop DPWS on sensors is how we can fit the IPv6 protocol into the sensor's memory and route IPv6 packets over energy constrained nodes. Since we need our packets to be carried over an Internet connection, IP must be utilized. For that reason, the 6LoWPAN architecture is utilized. 6LoWPAN is a protocol definition to enable IPv6 packets to be carried on top of low power wireless networks, specifically IEEE 802.15.4. The IEEE 802.15.4 standard defines two layers of the OSI model, the physical layer (PHY) and the Media Access Control (MAC) layer. Furthermore, it uses the Ad Hoc On-Demand Distance Vector protocol (AODV) for routing packets, which is a reactive routing protocol appropriate for energy constrained nodes. In the transport layer, we adopt the use only of the UDP protocol for all the interactions between the devices in order to minimize network traffic overhead. UDP removes features such as recovery of lost data and pre allocation of network resources and so the network load is reduced. Moreover it requires no connection establishment which can add delay, and it has a small segment header. We eliminate the use of SOAP and

HTTP protocols. As they are text-based protocols, they come with a certain amount of overhead which consumes network bandwidth and processor time. SOAP is a simple XML-based protocol that lets applications exchange information over HTTP. Consequently, HTTP using XML format for the transferring of messages constitutes a SOAP implementation. Our proposal is based on the *application specific format technique*, which constitutes the proposed technique for reducing the XML message size without prohibiting the usage of SOAP and other Web services standards. This technique allows the wireless sensors that host Web services and the component of the enterprise system which constitute the clients that may use the Web services hosted by the wireless sensors to know in advance the appropriate format of the XML document that they are going to exchange, due to the application specific type of data. Therefore, there is no need for a full SOAP implementation in the wireless sensor, as the compressed SOA messages can be manipulated and understood directly by the sensor, which is designed to have that ability. In this paper, we call compressed SOA message every message that is produced according to the *application specific format technique*. At the application level of the SPWS we do not consider the implementation of the entire HTTP server as we utilize only the HEAD and response messages of the HTTP protocol. In this way RAM is saved. This allows the invocation of a Web service to be as simple as sending an HTTP HEAD and response messages. Since in the proposed approach the application specific format technique is used, there is no need for the wireless sensor to use XML Schema for viewing a document at a relatively high level of abstraction and therefore considering its validation. As far as WSDL 1.1 is concerned, a modified version of this XML-based language is used for describing the wireless sensor's Web services and how to access them. The predefined knowledge acquired by the application specific format technique is utilized here too. The only information that the client and the wireless sensor need to exchange via WSDL is the data types used by the Web services that are hosted by the wireless sensor, as the wireless sensor can sent various types of data (for example temperature, pressure, speed etc.) depending on the requested by the client Web service. Obviously, this modification of the WSDL requires less RAM and fewer exchanges of messages. Instead of implementing the whole WS-Addressing which is a specification that allows Web services to communicate addressing information with transport-neutral mechanisms, a 'tiny' WS-Addressing protocol is used. WS Addressing protocol defines two interoperable constructs, the endpoint references and the message information headers,

VI. MODELING SENSORS AS SERVICES THROUGH SENSOR PROFILES FOR WEB SERVICES

1) The middleware architecture

The overall architecture of the proposed system based on the proposed middleware that converts a WSN to a Web services SOA Network connected to the Internet is depicted

in Figure 7. The wireless SPWS sensor is the wireless sensor that conforms to the SPWS while the Extended SPWS client is every client that conforms to the traditional DPWS and has also the mechanism that transforms compressed SOA messages which are based on sensor profiles for web services to complete SOAP/XML messages understood by the traditional DPWS and vice versa. This mechanism may be loaded on the client that is considered to have significant computing power and memory. Every wireless sensor sends, receives and handles compressed SOA messages. This allows significant savings in bandwidth. The wireless channel implies communication limitations, such as channel bandwidth constraints, time-varying fading, low QoS, etc. Therefore it is of a very importance to alleviate the traffic load with which wireless sensors burden the wireless channel. Furthermore, due to the fact that we need low-rate and low-cost wireless personal area network in order to make savings in battery consumption of sensors, the IEEE 802.15.4 standard is utilized for the communication between the wireless sensors. The basic framework of this standard conceives a 10-meter communications area with a transfer rate of 250 kilobit/sec.

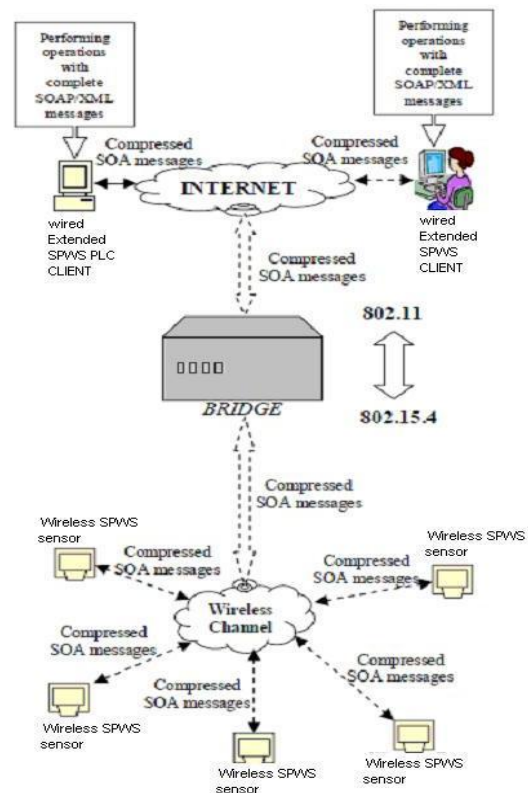


Figure7.The middleware using SPWS messages

VII. EXTENSIBLE CLOUD ARCHITECTURAL MODEL

The proposed paper aimed at working on applying and extending the service oriented paradigm to sensor network application engineering such as distributed manufacturing, we derived the requirements for the sharing of sensor networks as new resources in this domain. The necessary abstraction was implemented using the service oriented process parameters, which lead to the intelligence integration into the Internet. This solution has been extended to sensor clouds, which leads to high availability and hence reliability is achieved. This architecture presents the Integration Controller and Internet can interact using cloud technology. Cloud computing is a way to increase capacity or add capabilities on the fly without investing in new infrastructure, training new personnel, or licensing new software. The proposed architecture Enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications. Supports complete sensor data life cycle from data collection to the backend decision support system. Vast amount of sensor data can be processed, analyzed, and stored using computational and storage resources of the cloud. Allows sharing of sensor resources by different users and applications under flexible usage scenarios. Enables sensor devices to handle specialized processing tasks. In this architecture the IC will upload the sensed data to STAX. Figure 8 shows our implementation of sensor information in stax proxy

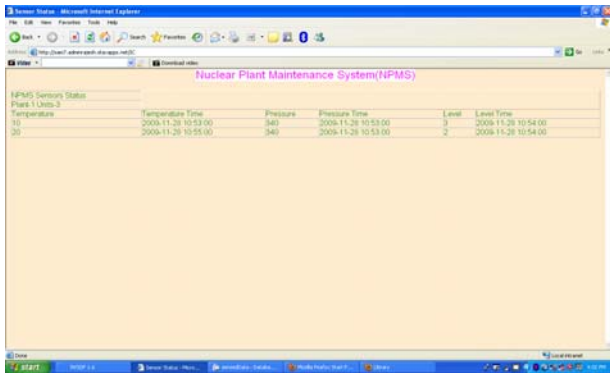


Figure 8. The deployment of temperature sensor as a service in cloud

VIII. PERFORMANCE ANALYSIS

The sensor network is simulated in TOSSIM with 300 nodes and the nodes are placed in grid topology. The message exchanged by the sensors with Sink/Gateway are

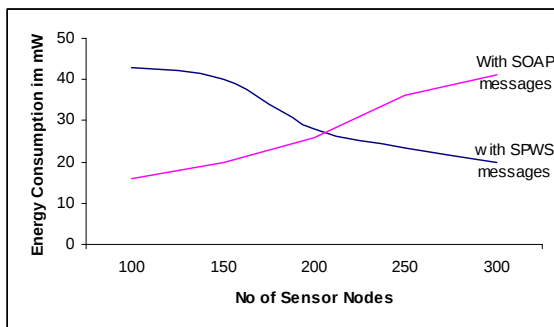


Figure 9. Energy consumption during message exchange

analyzed with traditional SOAP messages and the messages used with Sensor Profiles for Web Services. The energy consumption of sensor nodes is reduced compared to traditional SOAP messages. Similarly with increased number of message exchanges, the memory usages in sensor nodes are becoming constant with SPWS whereas it is increased with SOAP messages. The communication cost for sensors are more than their processing cost. Hence the reduction of power consumption and memory usage leads to less communication cost which enables an energy efficient model of sensor networks, which also increases the lifetime of sensor networks.

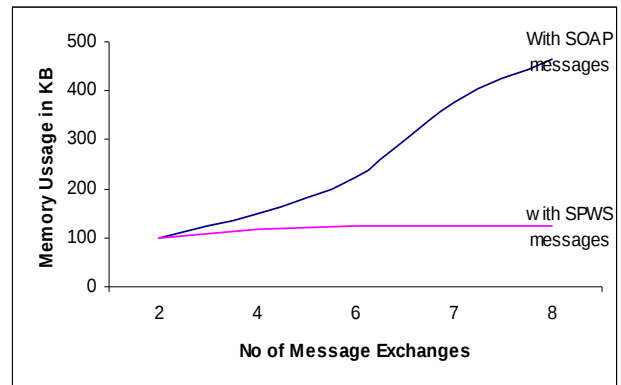


Figure 10. Memory usage of sensors

IX. CONCLUSION

In this paper a middleware using SPWS has been proposed. The lowest hierarchy level is considered to be a Wireless Sensor Network (WSN). The proposed advanced middleware implements a stripped down version of the traditional DPWS protocol. It was shown how a WSN that implements the proposed middleware can be integrated into the information system of an enterprise at a high abstraction level. Moreover, a technique was utilized which compresses and reduces the data volume of XML documents at a level that can be handled by the wireless sensors. The use of the SPWS in conjunction with this technique results to savings in the battery life of sensors, in reducing the memory requirements when storing XML exchanged documents and in reducing the traffic with which the sensors load the wireless channel. A novel way of combining wireless sensor networks with cloud computing services is presented. The wireless sensor networks are used to sense and collect environmental data. Since wireless sensor networks are limited in their processing power, battery life and communication speed, cloud computing usually offers the necessary storage capacity and processing power for long term observations, analysis and use in different kind of environments and projects.

REFERENCES

- [1] Flavia Coimbra, Delicato, Paul.O F. Pires, "A Service Approach for Architecting Application Independent Wireless Sensor Networks", Cluster Computing 8, 211-221, Springer Science, 2005.
- [2] Ioakeim K. Samaras, John V. Gialelis, George D. Hassapis, "Integrating Wireless Sensor Networks into Enterprise Information Systems by Using Web Services", Third International Conference on Sensor Technologies and Applications, 2009.
- [3] Nissanka B. Priyantha, Aman Kansal, Michel Goraczko, and Feng Zhao, "Tiny Web Services: Design and Implementation of Interoperable and Evolvable Sensor Networks", SenSys'08, November 5-7, 2008, Raleigh, North Carolina, USA. Copyright 2008 ACM 9781595939906/08/11.
- [4] Nils Hoeller, Christoph Reinke, Jana Neumann, Sven Groppe, Daniel Boeckmann, Volker Linnemann, "Efficient XML Usage within Wireless Sensor Networks", WICON'08, November 17-19, 2008.
- [5] Guido Moritz, Elmar Zeeb, Steffen Prüter, Frank Golatowski, Dirk Timmermann, Regina Stol, "Devices Profile for Web Services in Wireless Sensor Networks: Adaptations and Enhancements", IEEE 2009.
- [6] Weimin Zheng, Pengzhi Xu, Xiaomeng Huang, NuoWu, "Design a cloud storage platform for pervasive computing environments", Cluster computing, 15th November 2009.
- [7] Werner Kurschl, Wolfgang Beer, "Combining Cloud Computing and Wireless Sensor Networks", Proceedings of iiWAS2009.
- [8] Arnon Rosenthal, Peter Mork, Maya Hao Li, Jean Stanford, David Koester, Patti Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing", Journal of Biomedical Informatics 43, 432-435, 2010.
- [9] Åke Östmark, Jens Eliasson, Per Lindgren "An Infrastructure for Service Oriented Sensor Networks", Journal of Computers, Vol. 1, No. 5, August 2006.
- [10] Bobek, A., et. al, "Device and service templates for the Devices Profile for Web Services", In: 6th IEEE International Conference on Industrial Informatics (INDIN2008), pp.797-801. Daejeon, 2008.

AUTHORS PROFILE

R.S.Ponmagal is working as Asst Professor in the department of CSE, Anna University of Technology, Tiruchirappalli. She has more than a decade of experience in Engineering Colleges and published papers in International conferences and journals. Her area of research interest includes wireless sensor networks, mobile computing and web services.

Dr.J.Raja is working as Professor in the department of ECE, Anna University of Technology, Tiruchirappalli. He has more than two decades of experience in Engineering Colleges and published papers in International conferences and journals. His area of research interest includes wireless networks, ATM networks, mobile computing and web services.

Computer Simulation tool for Learning Brownian Motion

Dr Anwar Pasha Abdul Gafoor Deshmukh

Lecturer, College of Computer and Information Technology, University of Tabuk , Tabuk., Saudi Arabia.
mdanwardmk@yahoo.com

Abstract— With advancement of computers and Internet, there has been an increasing trend in utilizing computers and Internet for teaching and Education. Advanced topics and High Tech experiments demand for lot of infrastructure. There are areas in living sciences and other subjects where the magnitude of information is so huge that it is more convenient to do the experiment by simulation using computers. The present work relates to development of computer based learning resource that can be used as an online teaching aid or an offline tool to study the Brownian motion in two dimensions. It is an interactive tool where the learner can change the parameters of interest to study the Brownian motion and actually see the effect by way of simulated graphics. Quantitative values of the parameters arising from the simulation study are also available to the learner. The Brownian motion is a phenomenon occurring at microscopic scale and thus actual study of the experiment requires real time monitoring of microscopic development and progress of the Brownian motion which is a tedious task and very difficult to implement in laboratory. To this effect, to carry out the experiment under different desired conditions, a computer based tool is developed that enables the learner to conduct experiment using simulation and see the results quantitatively. It is being uploaded on internet as an online e-learning material.

Keywords-Computer Simulation; CBT; Brownian Motion

INTRODUCTION

The world is witnessing a change in every aspect of life since last 2 decades. The advance in Information Technology has opened up new vistas in the history of education also. Computer has become a useful teaching aid. It has infact became an integral part of learning

E-LEARNING AND CBT

The term E-Learning means electronic Learning and it is basically the online delivery of information, Communication, training and learning. E-Learning seems to have a multiplicity of definitions to each of its users and the term seems to mean something different. A very comprehensive definition has been given by the Cisco System, which defines 'E-Learning is Internet-enabled learning. Components can include content delivery in multiple formats, management of the learning experience, and a networked community of learners, content developers and experts. E-Learning provides faster learning at reduced costs, increased access to learning, and clear accountability for all participants in the learning process.

The World Wide Web now offers the possibility of conducting CBT on a global scale, without the usual restrictions of

platform dependence, the cost of mailing materials, and identifying interested users. Instead of searching for people who would like to use a given CBT model, the Web allows the reader to find the material and review it at his own pace. Because Web pages can be updated at any point, it is possible to keep the CBT materials up to date without having to do anything except edit the HTML documents involved. It is also possible to keep channels of communication open between author and reader, even if the two have never met before or had any sort of contact. Finally, the use of server-side scripts and interactive forms make it possible for the reader to be tested on what he has learned without any need for supervision.

Computer-based training (CBT) services are useful where a student learns by executing special training programs on a computer relating to their occupation. CBT is especially effective for training people to use computer applications because the CBT program can be integrated with the applications so that students can practice using the application as they learn.

BROWNIAN MOTION

The Brownian motion has been named in the honor of the Scottish botanist Robert Brown. He has presented the theory of Random movement of particles. The Brownian motion is the random movement of particles discovered by the Scottish botanist Robert Brown has made microscopically observations in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies, which can be found in the miscellaneous botanical works of Robert Brown, Volume 1. In his experiment he found that "the pollen particles appeared to move in a completely random fashion".

The major contribution to the explanation, interpretation and application came from Albert Einstein in 1905 and Marian Smoluchowski obtained that "If Kinetic theory of molecules was right, then the molecules in the water would move at random, the small particles to move in exactly the same way described by Brown." This work of Einstein was further refined by Norbart Weiner a mathematician at MIT, in 1923, established the modern mathematical framework of Brownian motion.

OUTLINE OF ACTUAL WORK

The principles and logic of the Brownian motion is derived from standard literature such as books and research journals.

For the implementation of the CBT, necessary algorithms are designed and tested initially in C++ and later on implemented on a Windows based platform using Html java with strong graphics support for user friendly interactive input and output facility. This CBT tool simulates the Brownian motion with the parameters provided by the user. As the Brownian motion precedes the random movement of the particle performing Brownian motion is displayed graphically. During its motion the path traced by the movement is also displayed graphically. The basic concept involves the displacement of the particle performing the Brownian motion so the numerical values of the parameters of interest are separately recorded. At the end of the motion the variables of interest will be displayed and the user can go for a fresh experiment again or quit the program.

TESTING AND PRESENTATION

The opening page of the CBT is provided with active links to the resources within the CBT to make different features and information available. In addition to the links provided on the left giving information of related topics. The learner can visualize graphically and perceive the way the movement of the gas molecules behaves under changing conditions.

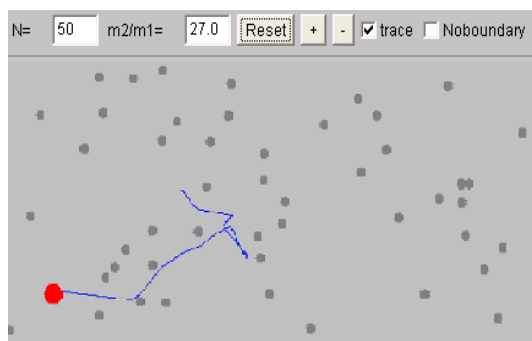


Figure-1(a) Brownian motion of 50 gas molecules'

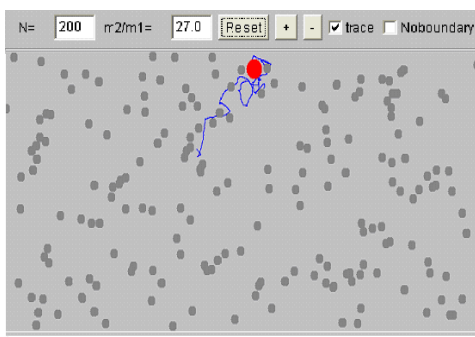


Figure-1(b) Brownian motion of 200 gas molecules'

To study the effect of pressure, one has to change the number of particles present in the container. As an increase in pressure causes an increase in density of the gas molecules, an increase

in pressure of gas results in an increase in the number of molecules present in a given volume of gas. By default the simulation window uses 100 molecules, an increase in value N corresponds to an increase in pressure (Figure-1(b)) and decrease in N correspond to decrease in pressure (Figure-1(a)). Thus the simulation of the movement of the gas molecules can be studied under changing pressure conditions. Random zigzag motion of the molecules when collide with each other undergo deflection and the path is changed that can be clearly seen in the simulation. The effect of change in pressure can be visualized clearly from the simulation.

In case if the particles collide with a particle with a different mass, the displacement of the particle will depend on the mass of the particle. To study the effect of the relative masses the ratio of $M1: M2$ can be changed; the value of $M1/M2$ is assigned a value of 27 by default This means the mass of the red molecule ($M1$) is 27 times greater than the rest of the molecules. This simulation allows for the study of the effect of $M1/M2$ on the distance moved by the particle in one collision or the mean free path when taken on an average basis. It can be clearly seen in the simulation that as the ratio of $M1/M2$ is large (Figure-1(c)), the displacement of the particle is small and for smaller values of $M1/M2$, this distance moved in one collision is large. Thus the process of Brownian motion can be clearly studied, perceived and understood by a learner which is not that easy in actual laboratory. This also demonstrates the concept of momentum transfer as the larger mass undergoes smaller displacement and vice versa. (Figure-1(d))

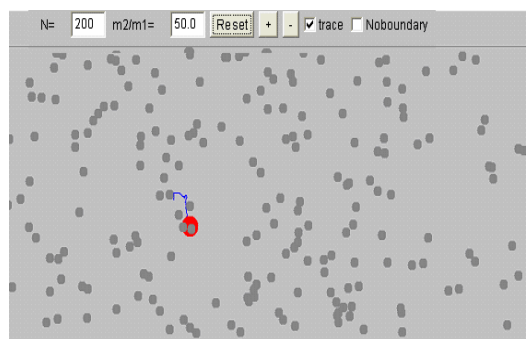


Figure-1(c) Brownian motion with 50 mass($m2/m1$)

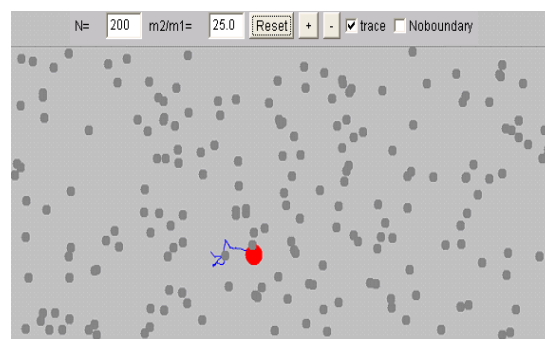


Figure-1(d) Brownian motion with 25 mass($m2/m1$)

The zigzag random motion of the molecules also depends on the temperature of the gas, to study the effect of temperature of the gas on the nature of Brownian motion, the simulation window is provided with two buttons with + and – sign to vary the speed of motion of the gas molecules. This speed of the gas molecules is in fact due to the kinetic energy of the

$$KE = \frac{1}{2}mv^2$$

molecules, where m is the mass and v is the velocity of the molecule. Thus increase in velocity of the gas molecules corresponds to increase in temperature and a decrease in velocity corresponds to a decrease in temperature of the gas. Thus With the help of the two buttons, the effect of change in temperature of the gas on the Brownian motion can be demonstrated. The learner can easily visualize the effect of the change in temperature on the Brownian motion and the effect of resulting collisions.

The two options Trace and No Boundary allow to choose the display of the path of the particle to be enabled or disabled, when check the trace option, the path traced by the red molecule remains visible as a blue trace of zigzag line else the zigzag line will not appear (Figure-1(e)). No boundary option makes the walls of the container as penetrable as if the medium (gas) is continuous and particle crossing one boundary appears from the other side so that the total number of molecules in the container remains constant.(Figure-1(f))

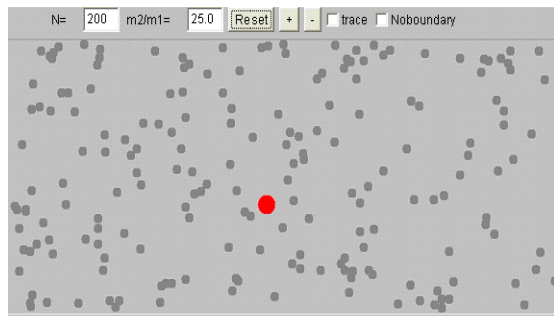


Figure-1(e) Brownian motion without Trace the path

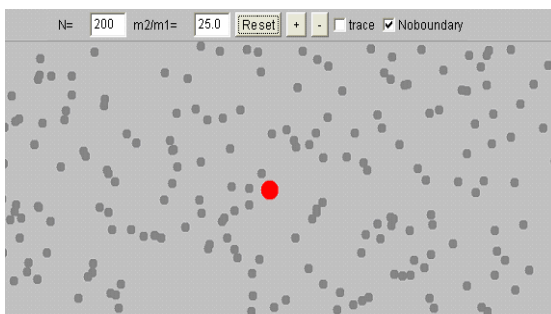


Figure-1(f) Brownian motion without Trace and
With Non boundary

Thus the role of the main parameters, gas pressure, temperature and the relative mass of molecules can

successfully be demonstrated through the CBT as described above.

CONCLUSIONS AND RESULTS

The CBT developed under the present project work has been successfully designed, developed, implemented and thoroughly tested. We used this CBT in a small class of students and found it to be very effective. It is being uploaded on Web to study the Brownian motion.

The CBT based on the study of Brownian motion was planned, designed and developed using techniques and tools described earlier. All the branching and links were carefully checked and tested for correctness.

REFERENCES

- [1] Robert Brown, August 1827 "A brief account of microscopical observations" Volume 1.
- [2] Stachel, John, 1998, "Einstein's Miraculous Year: Five Papers that Changed the Face of Physics" Princeton: Princeton University Press, p173
- [3] Jerison, David and Daniel Stroock, Number 4 (1995) , "Norbert Weiner." Notices of the AMS Volume 42, p432.
- [4] Chetan Srivastava, "Fundamentals of Information Technology", Kalyani Publishers, New Delhi.
- [5] Rajaraman V, 1996, "Fundamentals of Computer", (2nd Edition), New Delhi, Prentice Hall of India, ,
- [6] "Electronic learning - Wikipedia, the free encyclopedia" Computer Based Learning, sometimes abbreviated CBL, refers to the use of ... reuse of electronically-based teaching materials and in partucular creating en.wikipedia.org/wiki/E-learning

AUTHORS PROFILE

Personal Detail

Name : Deshmukh Anwar Pasha
Father Name : Deshmukh Abdul Gafoor
Date of Birth : 25-1-1977
Marital Status : Married
Languages Known : English, Marathi, Hindi, Urdu
Permanent Address : Near Aman Masjid, Motiwala Nager, New Baigipura, Indira Nager, Gali No 21, Aurangabad, Phone No. +966-544354254(KSU), +91-9326966582(India)
E-Mail : mdanwardmk@yahoo.com,
Educational Qualification

Doctor of Philosophy (Computer Science)
July-2010 Magadh University, Bodhgaya, India.
Master of Philosophy (Information Technology)
August-2008 Yashvantrao Chavan Maharashtra Open University, Nasik, India.
Master of Science (Information Technology)
June 2003 Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

Mobility Mnagement Techniques To Improve QoS In Mobile Networks Using Intelligent Agent Decision Making Protocol

Selvan.C

Department of Computer Science and Engineering
Government College of Technology,
Coimbatore, Tamil Nadu, India
selvantrichy@yahoo.com

Dr. R.Shanmugalakshmi

Department of Computer Science and Engineering
Government College of Technology,
Coimbatore, Tamil Nadu, India
shanmuga_lakshmi@yahoo.co.in

Abstract— Mobility Management (MM) is one of the major issues of Mobile networks that should be taken into account for providing Quality of Service (QoS) and to meet the subscribers demand (satisfaction). Mobility management techniques are divided into two types (i) Location Management (LM) and (ii) Handoff Management (HM). The LM is used for tracking where the subscribers are locate, and based on that permitting calls, Short Message Service (SMS) and other mobile phone services are delivered to them with the assistance of Intelligent Agent Decision Making Protocol (IADMP). The Redundant Update Remove Algorithm (RURA) which is used for reducing the updation between BSC and MSC. Enhanced Temporal Updation (ETU) reduces the location updates between Mobile Node (MN) and MSC using IADMP. To provide ubiquitous communication for subscribers without any break in the communication, HM protocols play main role in Mobile Networks. In HM process there are four protocols, (i) Double Threshold Protocol (DTP) (ii) Better Base Station Selection Protocol (BBSSP) using Relative Signal Strength (RSS) (iii) IADMP and (iv) Hybrid Decision Making Protocol (HDMP) are used. The Performances of these protocols are analysed. The HDMP and ETU provide QoS in Mobile Networks.

Keywords-component; Location Management;Handoff Mangement; Intelligent Agent; Enhanced Temporal Updation; RUR Algorithm; IADMP

I. INTRODUCTION

Mobility Management (MM) [3] is one of the major functions of a GSM or other Network. The aims of MM are to track where the subscribers are so that calls can be sent to them, and to record the subscriber services.

Mobile networks to provide quality of service (QoS) is challenging for the service providers. By introducing Intelligent Agent in MM it is possible to meet the QoS. The MM is divided into two major divisions as Location Management (LM) and Handoff Management (HM). The various MM techniques used in mobile networks are shown in Figure 1.

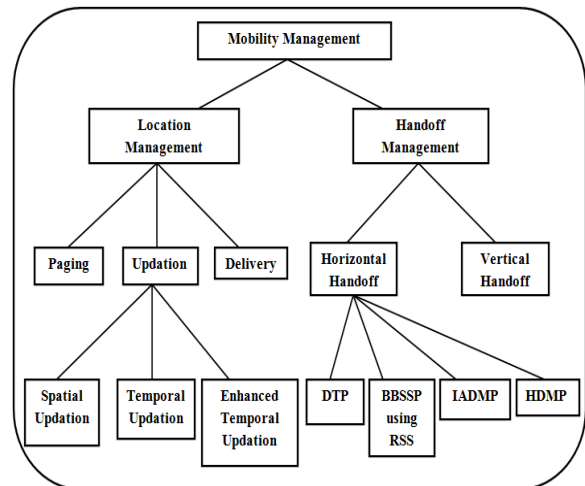


Figure 1. Overview of Mobility Management Techniques

In Location Management [5], the ability to manage information about the current location of mobile nodes based on their last update is a significant issue. That enables the network to discover the current attachment point of the mobile user for call delivery. Location Management is further divided into (1) Paging and (2) Call delivery and (3) Location updation.

An important issue in the design of Mobile network is how to manage the locations of mobile nodes and giving continuous connections to the subscribers wherever they go. The Mobile networks encompass the mobility management to provide ubiquitous and continuous communication to the subscribers, billing for the usage and further use.

An important issue in the design of Mobile network is how to manage the locations of mobile nodes and giving continuous connections to the subscribers wherever they go. The Mobile networks encompass the mobility management to provide ubiquitous and continuous communication to the subscribers, billing for the usage and further use.

II. LOCATION MANAGEMENT

The Location Management (LM) maintains the location of the Mobile Nodes to provide services. A micro cell [15] maximum coverage area is called location of a mobile node. A group of locations is called a location area. The LM is divided into three functions, updation follows paging and call delivery. The updating function makes the Mobile Node to update its place to the corresponding BSC. Paging operation is performed by the BSC to track all the locations of the mobile nodes at a time and periodically.

A. Analysis of Location Area

A GSM (Global System for Mobile Communication) network is a radio network of individual cells, called as base stations. Each base station covers a small geographical area which is division of a uniquely identified location area. By integrating the coverage of each of these base stations, a cellular network provides radio coverage over a much wider area. A group of base stations is named as location area.



Figure 2. Geographical location area

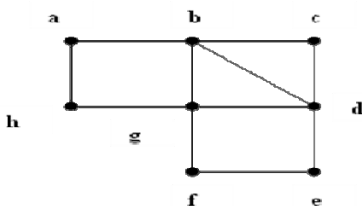


Figure 3. Geometrical location area

Figure 2, shows the geographical representation of location area and its divisions and the geometrical representation of location area were shown in Figure 3.

B. Paging

Paging is one of the fundamental mobility management procedures of a GSM network and also other cellular networks. MSC will request the BSC to scan all the active nodes under its coverage [1]. The BSC will send all the data to MSC after scanning all the nodes.

The Visitor Location Register (VLR) which normally knows the current location of the subscriber to the level of a location area. The VLR also knows which BSC controls cells in that location area. The VLR sends a paging request message to each of the relevant BSC. The BSC then send a paging request to every single cell within the location area. This paging request is broadcast on the cell broadcast channel to which the mobile is listening.

C. Call Delivery

When a mobile node tries to communicate with other mobile node, the call reaches MSC first through BSC, after that only the connection will be established to concern mobile node through BSC. Whenever the call comes from same or other network, the connection is established based on the Temporary Mobile Subscriber Identity (TMSI). TMSI is the identity most commonly sent between the mobile node and the network. An important use of the TMSI is in paging of the mobile.

D. Updation

The Location update procedure allows a mobile node to inform its active state to MSC through BSC. The updation is divided into three (1) Spatial updation (2) Temporal updation and (3) Enhanced Temporal Updation. This updation will be stored into the MSC through BSC. This updation process will take place periodically. The mobile node sends a message (location update request) to the network about its current location.

1) Spatial Updation:

The Spatial updation allows a Mobile Node (MN) to inform to the cellular network whenever it moves from one location area to the next. The mobile nodes should send updation to base station for receiving calls if any comes. When the mobile node does not send the updation it might be considered as a not reachable MN or switched off MN.

2) Temporal Updation:

The Temporal updation is performed based on the periodic time. The setting of interval time will be controlled by MSC. In the existing system the MN will inform the location of its exact place periodically. The location of the mobile node will be carried out to MSC through BSC.

Figure 4 gives the detailed operation of the existing update of the temporal updation. In this Figure 4, MN sends LU (Location Update) signal to BSC, the BSC forwards signal to MSC for updation. At the same time BSC sends Acknowledgement (ACK) to MN. After receiving signal from BSC, the MSC sends Location Acknowledgement (LA) to BSC. At last the BSC forwards the LA signal to MN. T_{LU} is time taken for single updation between MN and MSC. The operations are performed regularly by the mobile network. In this method, the update function takes place whenever the MN

sends update signal. Even though the mobile node is in the same location cell, the BSC forwards the update signal to MSC when it uses temporal updation. To solve this issue, the RURA (Redundant Update Remove Algorithm) is used. This algorithm removes the unnecessary updation between BSC and MSC.

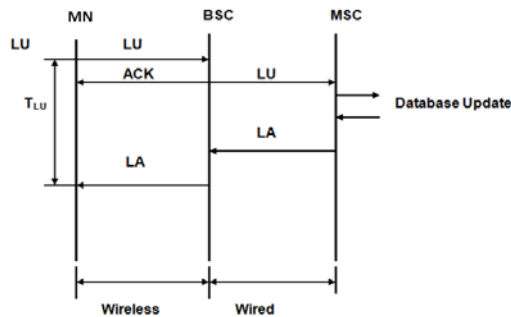


Figure 4. Existing System - Location Update

a) RURA (Redundant Update Remove Algorithm)

RURA removes the redundant update when the mobile node updates the same location. Figure 5, describes the functions of the mobile network using RURA. The BSC encompasses the RURA which never allow the same update to MSC. The Location Updation (LU) signal from MN and BSC, ACK signal from MSC and BSC are shown in figure 5.

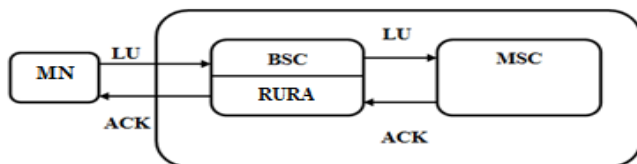


Figure 5. Functions of Updating Location using RURA

Whenever the location update takes place, the BSC will check the location symbol with existing buffer value, suppose the transmitted symbol not equal to existing symbol, it will update into the MSC otherwise it will not inform to the MSC.

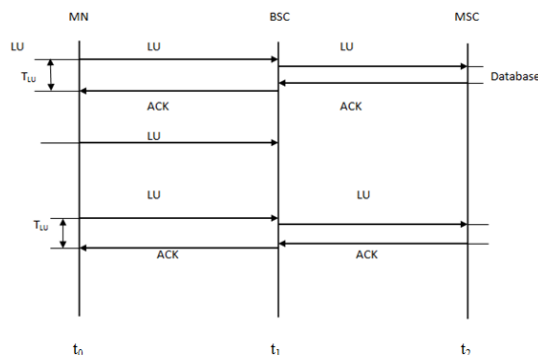


Figure 6. Timing diagram for RURA

In figure 6 to – Starting time for updation in Mobile node, t_1 – Time taken to reach BSC, t_2 – Time taken to reach MSC, T_t

$= t_0 + t_1 + t_2$, T_t - Total time for reaching MSC, T_{LU} - Time taken for location updation with ACK, The update signal is sent to BSC from MN in t_1 time. The update signal is sent from BSC to MSC in t_2 time. When the redundant signal comes from the Mobile node, the BSC will not respond to MSC and Mobile node. But it will update its buffer with time stamp.

The TABLE I. shows the movement records [7] of a single mobile node, for three hours using the temporal updation. The updation takes place every 10 minutes once. It updated 21 times for three hours. The Table II shows the movement records of a single mobile node using RURA; it needs only six updates for three hours. The updation takes place every 10 minutes once.

TABLE I. UPDATING RECORDS USING TEMPORAL UPDATION

Time Minutes-AM	Micro Cells																			
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
7.00	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.50	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.00	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.20	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.40	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.50	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.00	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.10	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.20	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.30	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.40	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.50	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10.00	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE II. UPDATING RECORDS USING RURA

Time Minutes-AM	Micro Cells																			
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
7.00	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.20	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.30	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.40	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

```
/* RURA Algorithm*/  
  
/* set timer=600 seconds*/  
  
label1:  
  
if(timer == 600)  
  
if(MS-signal != BSC-buffer)  
  
{  
  
MSC-update=MS-signal;  
database=MSC-update;  
go to label1;  
}  
  
else  
  
BSC-buffer= MS-signal;  
go to Label1;
```

Figure 7. RURA Algorithms

Figure 7 explains the RURA algorithm. The MN signal is checked with existing BSC buffer, whether equal to or not. If the MSC signal is not equal to BSC buffer value, the MN signal will be stored into the BSC buffer as well as the MN signal will be forwarded to the MSC.

The proposed RUR Algorithm reduces the updates only between BSC and MSC. In order to reduce the updates in BSC and also in MSC, a new technique known as Enhanced Temporal Updation (ETU) is introduced.

3) Enhanced Temporal Updation:

The Enhanced Temporal Updation (ETU) is a temporal updation technique which is performed based on (1) RURA and (2) IADMP (Intelligent Agent Decision Making Protocol). The RURA removes the redundant location updates. The MSC encompasses IA which has an IADMP (Intelligent Agent Decision Making Protocol). This Protocol predicts the updation of the MN. When the MN updation has to take place, when the MN updation has not to take place and how long the updates are needed? And how long no need? For these questions are answerable based on the movement history and call history [16] of MN. The IA analyse the each MN movement history and call history using data mining [[6], [2], [11]] process. IADMA sends the signal to the MSC to send update-lock-temp signal to Mobile node. The MSC temporarily stops the Mobile node updates using update-lock-temp signal. This signal will be sent from MSC to Mobile node through BSC. This update-lock-temp signal will be released automatically after the fixed time. Whenever the update-lock-temp signal is released by the mobile node, it will be informed to MSC. This information also will be analysed by IADMP for future use. Hence the Combination of IADM Protocol and RUR Algorithm is known as Enhanced Temporal Updation (ETU).

The TABLE III shows the comparison of packet updation using different updation techniques. The number of updation is reduced in ETU compared to spatial updation and temporal updation.

TABLE III. COMPARISON OF DIFFERENT UPDATION TECHNIQUES

No. of MS	Updates		
	Spatial Update	Temporal Update	Enhanced TU
20	1204	2108	998
40	2180	3120	1804
60	3210	4056	2909
80	4213	5108	3906
100	5107	6101	4806
120	6205	7284	5908
140	7106	8103	6806

Figure 8 shows the graphical representation of packet updation using different updation techniques.

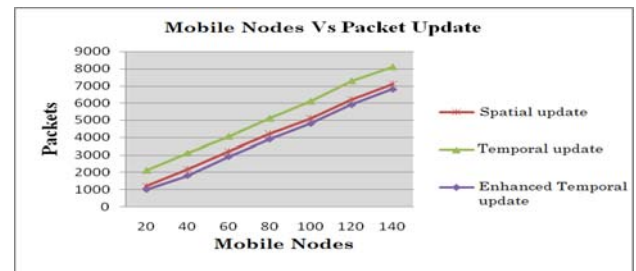


Figure 8. Mobile Nodes Vs Packet Update

TABLE IV shows the Comparison of delay in different updation techniques. Here the delay means, the time taken by the mobile node to receive the acknowledgement signal from MSC. The delay is reduced in ETU compared to other techniques.

TABLE IV. COMPARISON OF DELAY IN DIFFERENT UPDATION TECHNIQUES

No. of MS	Delay		
	Spatial Updation	Temporal Updation	ET Updation
20	0.45	1.54	0.36
40	0.48	1.59	0.38
60	0.59	1.64	0.42
80	0.63	1.69	0.46
100	0.69	1.73	0.42

Figure 9, a graph plotted for Mobile Nodes Vs Delay. The Enhanced Temporal updation technique compared with other techniques, the ETU takes less updation delay. In ETU the packet updation and delay are reduced compared to other techniques. Hence ETU technique provides QoS to Mobile networks.

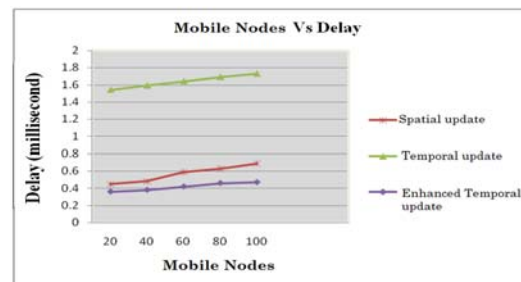


Figure 9. Mobile Nodes Vs Delay

III. HANDOFF MANAGEMENT

In Cellular Telecommunications, the term handoff refers to the process of transferring an ongoing call or data session from one channel to the core network. In Cellular communication handoff is the key matter to provide continuous service to the mobile nodes without any break. A single cell [8] does not cover the whole service area. There are two basic reasons for a handoff such as (1) The mobile node moves out of the range of cell (2) The received signal level decreases continuously until it fall below the minimal requirement for communication [12]. Hence Horizontal Handoff refers the same network handoff. The vertical Handoff refers the different network Handoff.

A. Analysis of Handoff Techniques

The Efficiency of the system handoffs [4] depends on the five characteristics (1) Minimum handoff latency (2) Low packet loss (3) Limited handoff failure (4) Intelligent Agent success rate high (5) Better Base Station Selection.

An analytical model has been previously developed for evaluating the performance of handoff algorithms based on Relative Signal Strength (RSS) [[4], [10]] measurements, i.e., the difference of signal strength from two Base stations, Absolute Signal Strength (ASS) which is the averaged value of the received signal level from current serving Base station measured by the mobile unit. The Measurement reports [8] including the quality of current link are sent to BSC by the Mobile node about every half second. The BSC receive the Mobile node signal and take decision, then it sends Handoff-request signal to MSC. Then MSC will send the signal to the BSC that activate the next Base station. Then the mobile node is taken over by the activated base station. The acknowledged signal is sent by mobile node to BSC and then BSC to MSC.

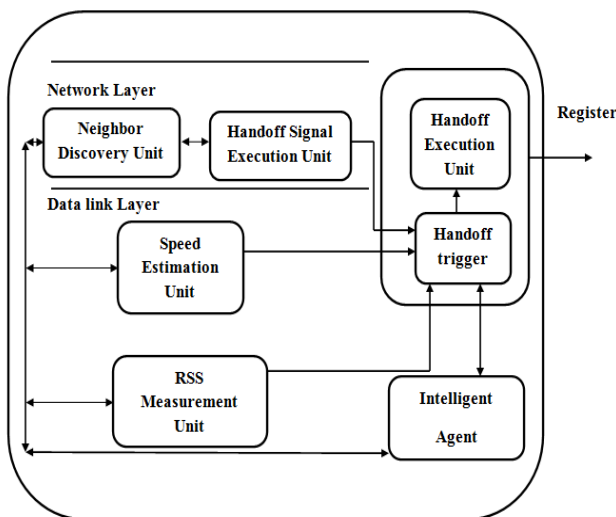


Figure 10. Proposed Handoff Architecture

As shown in Figure 10 the Proposed Handoff architecture [9] has seven modules. The Neighbour Discover unit discovers the neighbouring base station based on signal strength and

round-trip time. The speed estimation unit measures the speed of the mobile node. Based on the speed of the Mobile node and signal strength, the handoff is initiated in prior to support seamless communication. When the Mobile nodes move close to next base station, the base station informs to BSC about its very low signal. Then BSC takes decision and request the MSC for handoff. Many cases this kind of handoff initiations are failed. For example, the person carrying mobile node may go very close to Base station but he may change his direction at last minute. In this situation the handoff process may be wasted. In order to avoid this IADM Protocol is introduced to analyse the movement history of the Mobile node. At last IADMP will suggest whether handoff initiation should be done or not. Better base station is identified by BBSSP.

B. IADMP (Intelligent Agent Decision Making Protocol)

The Intelligent Agent Decision Making protocol is used to provide QoS in mobile networks. The IADMP does the following works i) Predicts the updation of the mobile node based on the mobile nodes service history and ii) Take the decision which base station is better when the handoff take place based on the mobile nodes handoff history. In this work Apriori Algorithm is used for mining frequent itemsets.

TABLE V. MOVEMENT HISTORY OF MOBILE NODE

Mobile Number: xxxxxxxxxx

Time	Location of MN							Day						
(7 am-8 am)	a	b	c	d	e	f	g	MO	TU	WE	TH	FR	SA	SU
7.10	1	0	0	0	0	1	0	0	0	0	1	1	1	1
7.20	1	0	0	1	0	0	0	0	1	0	0	0	0	0
7.30	1	0	0	0	0	0	0	1	0	0	0	0	0	0
7.40	0	1	0	0	0	0	0	1	0	1	0	0	0	0
7.50	0	1	0	0	1	0	0	1	0	0	0	0	0	0
8.00	0	0	1	0	0	0	0	1	0	0	0	0	0	0

The table V refers the data records of mobile node movements based on location history. In the initial process it collects all the details from MSC. The IADMP does the data mining process to find the frequent data sets using Apriori algorithm. The Apriori algorithm, the first step, it scans all of the transactions in order to count the number of occurrences of each item. Here number 2 is taken as a minimum support count. The algorithm counts the each candidate frequent occurrences, the frequency of the occurrences less than 2 means it will not consider for future use, otherwise it will be taken into account to take decision. At last the high frequently occurring pattern is retrieved. Using this pattern, decision is taken by the IADMP.

C. HDMP (Hybrid Decision Making Protocol)

The Hybrid DMP performs the functions based on the BBSS Protocol using RSS with IADMP information. The IADMP makes decision from the history of the Mobile node which is based on the data mining operations. The BBSS Protocol finds the better base station using signal strengths. Figure 11 shows the procedure for Hybrid DMP.


```

/*Hybrid DMP*/
initialization();
label1:
advertisement-from-AP();
MN-updation();
signal-strength-function();
if(signal strength != beacon)
call IADMP();
handoff-notice();
registration();
establish-IP-connectivity();
transfer-operational-parameters();
transfer-packets();
else
continue-communication();
call label1;

```

Figure 11. Hybrid DMP

The HDMP (Hybrid Decision Making Protocol) is a mixing of IADMP and BBSSP. The HDMP is compared with Double Threshold Protocol (DTP) and BBSSP. The simulation works are performed with different speed of mobile node and the results are tabulated in table VI.

TABLE VI. MEASUREMENT ANALYSIS OF MOBILE NODE

S. No	Measuring Signal strength and Time			
	Mobile node Speed (km/hour)	Distance between BS and MN (meter)	Signal strength (dBm)	Destination Time (millisecond)
1.	20	1162.4	-42.83	58120
2.	30	873.3	-40.34	29110
3.	40	612.8	-37.27	15320
4.	50	352.5	-32.46	7050
5.	60	258	-29.75	4300
6.	70	245	-29.30	3500
7.	80	196	-27.37	2450

Table VII shows the measurement analysis of Handoff. The handoff initiation takes place before Mobile node reaches the destination.

TABLE VII. MEASUREMENT ANALYSIS OF HANDOFF

S. No.	Mobile node Speed (km/hour)	Destination Time (millisecond)	Handoff Initiation (millisecond)	Signal strength for handoff initiation (dBm)
1.	20	30000	29990	-61.42089
2.	30	20000	19988	-61.30225
3.	40	15000	14984	-61.14403
4.	50	12000	11980	-60.94625
5.	60	10000	9974	-60.70892
6.	70	8571	8539	-60.43203
7.	80	7500	7460	-60.11559

Time Vs Signal Strength

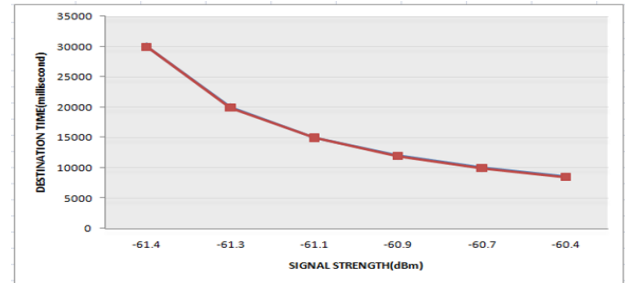


Figure 12. Destination time Vs Signal Strength

Figure 12 shows the graph for Destination time Vs Signal strength. When the mobile node moves to base station the destination time decreasing, simultaneously the signal strength of the mobile node increases.

TABLE VIII. COMPARISON OF DELAY IN DIFFERENT PROTOCOLS

Mobile node Speed(km/hour)	Delay(millisecond)		
	DTP	BBSSP	HDMP
20	1.21	1.02	0.9
30	1.34	1.21	1.13
40	1.40	1.31	1.19
50	1.51	1.35	1.24
60	1.56	1.42	1.30
70	2.12	1.80	1.35
80	2.15	2.03	1.48

The table VIII shows the comparison of different protocols, in this table delay refers to the time taken to send signal from Mobile node to MSC through BSC, as well as acknowledgement signal is sent back from Mobile node to correspondent BSC to remove old information. The Double Threshold Protocol has two threshold points to alert the handoff initiation. In this protocol the delay is high as shown in the table VIII. The Better Base Station Selection Protocol (BBSSP) selects the better base station for handoff using RSS (Relative Signal Strength) and ASS (Absolute Signal Strength). Hence the IADMP (Intelligent Agent Decision Making Protocol) takes less delay than BBSSP when handoff occurs. The HDMP (Hybrid Decision Making Protocol) is the combination of BBSSP and IADMP. The HDMP makes less delay than other protocols.

Speed Vs Delay

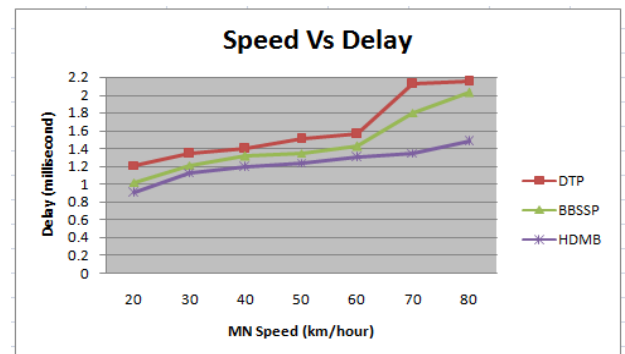


Figure 13. Graphs for Comparison of Handoff Protocols

Figure 13 shows the comparison of delay for different handoff Protocols. As shown in the graph HDMP takes less delay in Handoff.

IV. CONCLUSIONS AND FUTURE ENHANCEMENT

In mobility management, the following techniques are implemented i) RUR Algorithm ii) ETU (Enhanced Temporal Updation) technique iii) DTP (Double Threshold Protocol) iv) IADMP (Intelligent Decision Making Protocol) and v) HDMP (Hybrid Decision Making Protocol) using Network simulator 2.

A. Conclusion

Location management and Handoff management protocols are implemented and improved the Quality of Service in mobile networks. The proposed handoff architecture is designed to provide the continuous service to mobile user. The Service providers and subscribers are more benefited due to the technical advancement. Hence these protocols and algorithm can be used in the next generation wireless communications systems to reduce the location management signalling costs.

B. Future Enhancement

In coming years, IADMP will play major role in mobile computing. The Intelligent Agent decision making protocol will be adapted with existing protocols without making any difficulties. Of course, numerous protocols may be created to improve QoS but the requirement of the subscribers necessity may be varying based on the technological advancement. It may increase the software maintenance cost and system complexity, but there is no question for failure of the handoff failure.

ACKNOWLEDGMENT

We would like to express our cordial thanks to Dr. V. Lakshmi Prabha for her full support and valuable suggestions to continue our research work.

REFERENCES

- [1] Abhishek Roy, Archan Misra and Sajal K. Das (2007), "Location Update versus Paging Trade-Off in Cellular Networks: An Approach Based on Vector Quantization", Vol 6, pp. 1426 – 1440.
- [2] Abraham Silberschatz, Henry F.Korth and S.Sudarshan(2006), MC Graw Hill Highe Education publication, "Database System Concepts" pp. 723-753.
- [3] I.F. Akyildiz et al. (2007), "Mobility Management for Next Generation Wireless Systems" Proc. IEEE, Vol. 87, No. 8, pp. 1347-1384.
- [4] Ezil Sam Leni. A and Srivatsa. S. KH (2008), "A Handoff Technique to Improve TCP Performance in Next Generation Wireless Networks", Information Technology Journal 7(3). pp. 504-509.
- [5] Ian F.Akyildiz and Wenye Wang (2002), "A Dynamic Location Management Scheme for Next Generation Multitier PCS Systems", IEEE Transactions on Wireless communications, Vol. 1, No. 1.
- [6] Jiawei Han and Micheline Kamber (2001), "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers (An imprint of Elsevier).
- [7] Jin Soung Yoo, Shashi Shekhar. (2009), "Similarity-Profiled Temporal Association Mining" IEEE Transactions on Knowledge and Data Engineering.

- [8] Jochen H.Schiller(2006), "Mobile Communications" Pearson Education, Ltd.
- [9] Liebsch.M, A.Singh, H.Chaskar, D.Funato and E.Shim, (2004), "Candidate Access Router Discovery", Internet Draft, Internet Engineering Task Force, draft-ietf-seamoby-card-protocol-07.txt.
- [10] Ning Zhang and Jack M.Holtzman (1996), "Analysis of Handoff algorithms using absolute and Relative measurements" IEEE transactions on Vehicular Technology, Vol. 45, No. 1.
- [11] Patricia Serrano-Alvarado, Claudia Roncancio, Michel Adiba, (2004), "A Survey of Mobile Transactions", Source Distributed and Parallel Databases archive Vol. 16 , Issue 2 .
- [12] Shantidev Mohanty and Ian F.Akyildiz, (2006), "A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next-Generation Wireless Systems", IEEE Transactions.
- [13] Shiow-yang Wu and Hsiu-Hao Fan (2010), "Activity-Based Proactive Data Management in Mobile Environments" IEEE Transactions on Mobile Computing, Vol. 9, No. 3.
- [14] The Ns Manual <http://www.isi.edu/nsnam/ns/nsdocumentation.Html>
- [15] Theodore S. Rappaport (2007), "Wireless Communications Principles and Practice" Prentice-Hall of India Private Ltd. New Delhi.
- [16] Xian Wang, Pingzhi Fan, Jie Li and Yi Pan, (2008) Modeling and Cost Analysis of Movement-Based Location Management for PCS Networks With HLR/VLR Architecture, General Location Area and Cell Residence Time Distributions", IEEE Transactions on Vehicular Technology, Vol. 57, No. 6.

AUTHORS PROFILE



Selvan.C is doing his research in the area of Mobile Computing in the Department of Computer Science and Engineering in Government College of Technology, Coimbatore, Tamil Nadu, India. His research work is supported by University Grant Commission, New Delhi, India. He has registered his research in Anna University of Technology, Coimbatore, Tamil Nadu, India. His area of interest is Mobile Computing and Mobile Communication. He is the student member of IEEE.



Dr.R.Shanmugalakshmi is working as an Assistant Professor in the Department of Computer Science and Engineering in Government College of Technology, Coimbatore, Tamil Nadu, India. She has published more than 40 International / National Journals. Her research area includes Image processing, Neural Networks and Mobile Computing. She has received Vijaya Ratna Award from India International Friendship Society in the year of 1996. She has received Mahila Jyothi Award from Integrated Council for Socio-Economic Progress in the year 2001 and she has received Eminent Educationalist Award from International Institute of Management, New Delhi in the year of 2008. She is a member of Computer Society of India, ISTE and FIE.

Security Risks and Modern Cyber Security Technologies for Corporate Networks

Wajeb Gharibi,
College of Computer Science and Information
Systems, Jazan University, Jazan, Saudi Arabia.
gharibi@jazanu.edu.sa

Abdulrahman Mirza,
Center of Excellence in Information Assurance
(CoEIA), King Saud University, KSA.
amirza@ksu.edu.sa

Abstract—This article aims to highlight current trends on the market of corporate antivirus solutions. Brief overview of modern security threats that can destroy IT environment is provided as well as a typical structure and features of antivirus suits for corporate users presented on the market. The general requirements for corporate products are determined according to the last report from av-comparatives.org [1]. The detailed analysis of new features is provided based on an overview of products available on the market nowadays. At the end, an enumeration of modern trends in antivirus industry for corporate users completes this article. Finally, the main goal of this article is to stress an attention about new trends suggested by AV vendors in their solutions in order to protect customers against newest security threats.

Index Terms—Antivirus technologies, corporate security, corporate network, malicious software, protection, threats, trojan.

I. INTRODUCTION

MOST companies think of defeating itself against potential security attacks, but only a few of them really imagine a set of security threats that can danger the company. Many of them described in corporate in security standards thus helping the companies to organize IT security defense system. In such context antivirus protection plays the vital part of whole security area. Moreover contemporary antivirus solutions become more advanced and mature. Nowadays they include not only antivirus engine for workstations and an administration console, but many additional features, like antivirus for a mail protection system, a gateway, a database of incidents and enhanced report and logging system. Nonetheless, an implementation of many of such solutions is far from solving all corporate security issues. That is why it is not enough to install only personal antivirus products within a corporate network, but whole corporate suite to cope with all threats at different levels of a network. This will help to construct a corporate secure IT environment.

II. THE RISK OF MALWARE AND INTERNET THREATS

The main risks for companies in area of information security comprise infections by viruses, trojans, worms, exploits and other malicious code that can reveal the corporate secrets by stealing confidential data and be the reason of serious data leakage. Also phishing and online banking fraud can be a serious problem for IS managers.

Taking in consideration that corporate IT infrastructure mainly consists of domain-joined computers it can be more likely to encounter worms. The main propagation vectors of worms are opened file shares, removable drives, e-mail and IM channels. These are commonly used within companies' networks as a corporate communication and can be a potential threat. According to Microsoft Security Intelligence Report [13], 4 of the top 10 malware families detected on domain-joined computers are worms.

The most popular families are Autorun worms that can spread through removable drives, and network worm Kido/Kido/Conficker/Downadup which was appeared on November 2008 and caused a global world epidemic. The worm has struck more than 10 million computers, using vulnerability in service "Server" (MS08-067).

The worm sent to the remote machine specially crafted RPC-request on TCP port 445 (MICROSOFT_DS[SMB]) which caused the buffer overflow by calling `wscpy_s()` function in `NetpwPathCanonicalize()` (library `netapi32.dll`). The given malicious program applied a wide spectrum of methods to hide the presence in the system: files view settings in Explorer, disabling the services, responsible for system security. It was used several ways of distribution: the admin shared folders, removable devices, downloading the updates from websites, domain addresses of which were generated by special algorithm. As a result it has received a wide proliferation all over the Internet. The detailed description of the worm you can find in malware encyclopedia [14].

* This work was partially supported by CoEIA, King Saudi University, KSA.

III. SECURITY RISKS IN HARDWARE

A. Overview

Recently Dell Company, the leading computer system manufacturer, announced that in its servers' line PowerEdge malicious program has been found embedded in a flash memory of a motherboard [15].

Thus, the computer industry has been faced with the threat of computers' infection with malicious software, but at the level of firmware. The topic of malicious inclusions in hardware is becoming more importance due to the fact that most of our systems on chips are fabricated in Southeast Asia, although under the brand names of major U.S. companies. This can be explained by reducing production costs and increasing market competitiveness. Another side of a coin is losing a trust during a fabricating process. Especially, when it comes to development for military purposes, which may result in decommissioning weapon systems.

A model of compromised system is represented in Fig. 1.

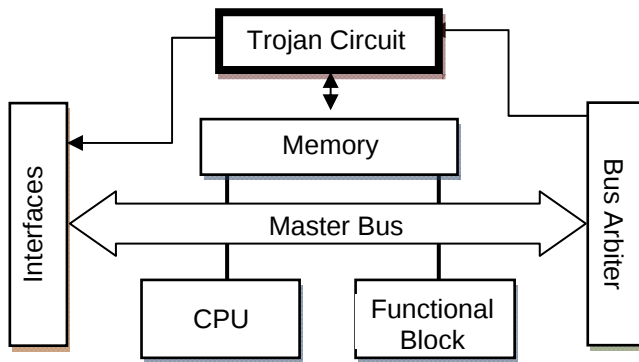


Fig. 1. Trojan insertion embedded in system on a chip

The trojan can be activated by a special value on Master Bus, for instance, it can be memory address where stored targeted data. Once trojan circuit is triggered, the payload can be one of the following: disabling system, transmitting interested data to third party by means of embedded interfaces, collecting accessed information in the memory for further utilization, rising security privileges for a current process running in the system.

B. A Formal Model of Hardware Trojan (HT)

Let us consider a formal model of HT by introducing several abstract concepts. *Trojan* (T_i) is a malicious component that can provide an access to *System* (S_i) in certain moment with the appropriate condition.

The pairs (T_i , O_i) are bound by the set of specified actions A_s . This set is defined according to security policy and specification of the vendor and is a subset of the whole set A of all possible actions for each pair.

At the same time, pairs (T_i , O_i) can communicate by a set of malicious actions A_m . It is obvious that $A = A_s \cup A_m$.

The purpose of security verification is identifying actions from the set A_m .

The task of malicious circuit detection is getting more complicated when HT can take advantage of a set of specified actions, that can gain an access to a computer system or its component, from the set A_s , such as $A_s \cap A_m \neq \emptyset$. As a result, it is needed to verify system considering a whole set of actions A . It is a hard verification task even for small systems on a chip because of searching within a set of all possible input vectors.

C. Hardware Trojan Detection Task

The danger hidden in complex system on a chip nowadays is underestimated. The trojan circuit can be easily embedded to a system on a chip and hardly detected taking in consideration the size of the modern digital system [18].

The formal view to the problem of malicious insertions proves that the task of trojan detection in complex digital system is difficult.

The solution can be found in the area of high level testing methodology in order to cope with the complexity of the task. Nowadays there are powerful methods that are provided by researchers that can help in trojan detection and analysis, such as in [19] and [20], but still there is no mature solution that can provide universal methodology for fables companies and governments.

IV. ASSESSING THE LOSSES OF THE COMPANY FROM SECURITY THREATS

The breaches in corporate environment may cause undesirable data leakage and will lead to suspending business processes of the company. In such scenario it may lose important customers and business partners because company which cannot protect itself from this attacks is faithless over the unforeseen costs like malicious programs influence, information drain, attacks on computer networks, etc [16].

The result is that when the number of personal computers is growing and communication channels capacity is increasing malware epidemic's scope and losses are growing correspondingly. Therefore the company management has to think about the information security.

In modern world the probability of malicious programs get into a computer system is constantly growing. It may cause not only short-term fault in the network, but a complete stopping the company. Losses by malicious programs are estimated as billions of dollars around the world annually and continue to increase.

According to [17] the cost of the average caused by malware attack in a corporate network can be calculated as in (1).

$$\begin{aligned} \text{DELAY} = & (\text{comp_num} \times \text{fix_time} \times \text{adjuster_hour_payment}) \\ & + \text{additional_expenses} + \\ & + \left(\frac{\text{items_day} \times \text{product_price} \times \text{fix_time} \times \text{comp_num}}{8 \times \text{adjuster_num}} \right) + \\ & + \left(\frac{\text{salary} \times \text{comp_num} \times \text{fix_time}}{8 \times 22 \times \text{adjuster_num}} \right), \end{aligned} \quad (1)$$

where *comp_num* – number of computers within a network; *fix_time* – time in hours for fixing a fault; *adjuster_hour_payment* – payment for adjusting a computer per hour; *adjuster_num* – number of such specialists; *additional_expenses* – additional expenses for network repairing and buying new devices; *product_price* – price of a product; *items_day* – number of product items per day; *salary* – salary of an employee per month.

V. ANALYSIS OF CORPORATE ANTIVIRUSES

According to latest report from Av-Comparatives Lab [1] the main players at corporate security market are Avira, Eset, G Data, Kaspersky, Sophos, and Symantec. In this article we will overview functional diversity of existed corporate suits and take a look to nearest future of corporate antivirus suits which seem to become a total security solution for corporate users.

The typical structure of corporate suite:

- 1) Administration console – provides useful managing and configuration environment for administrators of big networks.
- 2) Antivirus for workstation – actually the antivirus engine with all features peculiar to workstation antivirus. Provides centralized protection of user's system on a corporate network against all types of malware, network attacks, spam.
- 3) Mail server antivirus – protects the mail server against spam and malware delivered by email channels.
- 4) File server antivirus – protects data on servers under Microsoft Windows operation system control against all types of malware. Designed mostly for high-performance corporate servers.

Analyzing all products options it has been distinguished the main features of modern corporate antivirus:

- 1) Easy Installation and Deployment – simple and fast way to deploy the solution into a big corporate network, supporting Active Directory technology.
- 2) Usability and Management – console provides useful management interface with real-time monitoring and logging features.
- 3) Scalability – solution works with networks of different size from small business to enterprise scale with thousands of computers distributed geographically around many offices.
- 4) Technical Support and Updates – regularly delivers antivirus updates and helps to solve all unforeseen

security issues of a company with a short response time. Also website and online services are important points.

- 5) Cross-Platform Security – ability to protect systems with different types of operational systems, such as Linux, MacOS, mobile platforms, etc.

VI. NOWADAYS TRENDS AND SUGGESTIONS

As for future in area of corporate users' security protection a growing trend is including more sophisticated administration interface that provides detailed information about the real-time status of the network. It can be represented as advanced graphical interface with diagrams or even as a separate product. It can be an intelligent agent that can handle huge amount of information from thousands of computers and hints the administrator what to do in that case.

For instance, Blue Medora designed a special agent for Symantec corporate solution which results in "less complexity, more uniform operations management, and a significant reduction in costs due to the elimination of redundant infrastructure and multiple platform-specific tools" [2]. It proves the idea that there is an area for further improvement of corporate antivirus solution even for the outstanding vendor.

Among extensible features are the following:

- 1) Improved monitoring of incidents with malware.
- 2) Improved monitoring of the user's intrusion into the antivirus key processes.
- 3) Monitoring of failures in updates and malware scanning tasks.

The new features to be included into the product:

- 1) Real-time status and availability monitor.
- 2) Log monitors.
- 3) Report and take-action system that would help administrator to perform necessary actions to any type of threats.

The main idea is to raise a sensitivity level of the persons who are responsible for corporate network security and reduce the time of reaction to the emerging danger. Therefore, useful and exhausted data representation can really help in struggling against malware.

Except logging and monitoring an essential part of security solution is integrity. Modern corporate antivirus solutions comprise not only a bunch - Antivirus, Antispyware, Firewall, Antispam with Managing System, but many additional features, such as Backup systems, Password and Key Managers and Encryption Utilities to organize safe confidential data storage.

This trend is peculiar to home solutions as well. Thus, Kaspersky Pure for home users provides besides malware protection also password management system to keep in safe all family's identities [3]. Another example, Norton Online Family also allows observing the kids activity on

computer [4].

As for enterprise suits, Symantec provides Protection Suite for Endpoints where gathered encryption, confidential data storage and others features aimed to maintain IT security in a company [5]. One more interesting product came from Sophos [6]. Endpoint Security and Data Protection has Integrated DLP (Data Loss Prevention) and Encryption tools in its package.

Also mobile and non-Windows platforms should be supported within a corporate solution because of huge diversity of working devices: laptops, PDAs, smart phones, etc. Many antivirus vendors have such solutions in a product line.

The important point to be considered is Security-as-a-Service. A security is not only software, but a state of a system. It is important to have 24/7 technical support service to solve a newest security issues, such as new versions of malware, zero-day exploits. Often proactive defense cannot cope with a huge variety of new malware modification released every day by hacker's generators. The same way administrator cannot keep all software up-to-date with new patches installed. In such context deploying vulnerability searching system is desirable to reveal software breaches and notify to install new updates in time.

Here the problem of support service's quality has been raised. It is not a secret that a high quality support service can be granted only by the team of qualified malware experts not by "sandbox" robots [here we can put a reference to our research in "Sandbox Comparatives"]. Many companies provide malware analysis column on their web sites or even separate security domains where the descriptions of most popular threats are published, like it is done at virusradar.com by Eset and securelist.com by Kaspersky Lab.

Another side of the coin is an ability to remove consequences of an infection. Not all antivirus engines allow proper disinfection of the system or network after an incident that already has taken a place. In that case special removal utilities and scripts are released by analysts to help administrators in cleaning their IT farms. There are such services from Symantec [7] and AVZ tool from Kaspersky Lab [8].

Phishing is becoming a serious problem for all users in the cyber world. What antivirus vendors can suggest in protecting corporate users against this problem except of standard anti-phishing modules that block dangerous web sites from black list? The interesting solutions have been introduced within Kaspersky Internet Security 2011 – Geo Filter and Online Banking modules. According to information from official site: "Geo Filter provides the user with an option to block domains related to specific countries. Online Banking controls requests to Online Banking services while processing confidential data" [9]. Those modules could be helpful in keeping a communication with financial institutions more safe which could be essential in corporate environment.

Finally, a following to corporate security standards is what some AV vendors do. The big companies try to organize corporate IT security according to policies compliant to security standards. Among them:

- 1) X509 – is ITU-T standard specifies formats for public key certificates, certificate revocation lists, attribute certificates, and a certification path validation algorithm [10],
- 2) LDAP (Lightweight Directory Access Protocol) – is an application protocol for querying and modifying data using directory services running over TCP/IP [11],
- 3) Microsoft IWA (Integrated Windows Authentication) – provides authentication connections between Microsoft IIS, Internet Explorer, and other Active Directory aware applications [12].

VII. CONCLUSION

To sum up, in this short review the current security threats have been briefly presented. According to them an analysis of antivirus solution for corporate users was proposed. The general features and structure of corporate suit were enumerated based on the latest report from av-comparatives.org. In the last part of the article we considered modern trends in current antivirus solutions from most popular AV vendors, such as Eset, Symantec, Sophos and Kaspersky.

It is obvious that the corporate products represent quite powerful solutions for enterprise networks but they could become better by adopting new standards, technologies and a high level of support services. The corporate suit is becoming a heavy package of tools aimed to fight against malware, network attacks, spam, phishing. It gives to administrators a control under a huge corporate network that allows monitoring a real-time activity and react to an existing situation as soon as possible. A corporate security is a multifactor system that consists of security software, services, policies and a human factor. None of them should be missed in a building process of secure corporate environment.

REFERENCES

- [1] "Review of IT Security Suites for Corporate Users", May 2009. Available: www.av-comparatives.org.
- [2] Blue Medora Agent for Symantec Endpoint Protection, Available: <http://www.bluedora.com/product/page/40>
- [3] "Kaspersky Pure. Ultimate Protection for Your Digital Life", Available: <http://www.kaspersky.com/kaspersky-pure>
- [4] "Norton™ Online Family. A smarter way to keep your kids safe online". Available: <https://onlinefamily.norton.com/familysafety/loginStart.fs>
- [5] "Symantec Protection Suite Enterprise Edition for Endpoints". Available: <http://www.symantec.com/business/protection-suite-enterprise-edition-for-endpoints>
- [6] Sophos, "Endpoint Security and Data Protection". Available: <http://www.sophos.com/products/enterprise/endpoint/security-and-control/>
- [7] Symantec, "Spyware and Virus Removal". Available: <http://www.symantec.com/norton/nortonlive/spyware-virus-removal.jsp>

- [8] "How to scan your computer, save the log and run a script using the AVZ utility? ". Available:
<http://support.kaspersky.com/faq/?qid=208279710>
- [9] Kaspersky Internet Security 2011 Manual. Available:
<http://support.kaspersky.com/kis2011?level=2>
- [10] Wikipedia, <http://en.wikipedia.org/wiki/X.509>
- [11] Wikipedia, <http://en.wikipedia.org/wiki/LDAP>
- [12] Wikipedia,
http://en.wikipedia.org/wiki/Integrated_Windows_Authentication
- [13] Microsoft Security Intelligence Report. Volume8. July-Dec2009, Available: <http://www.microsoft.com/security/about/sir.aspx>
- [14] Malware Encyclopedia. Available:
http://www.totalmalwareinfo.com/eng/Net-Worm.Win32.Kido/_Conficker.A-C_Worm
- [15] "PC giant warns of hardware trojan", NewScientist, 22 July 2010.
- [16] Filatov, Kozlovskih, Cvetkova, Planning, Finance, Management of Enterprise. – Finance and Statistics, 2005, 384 p.
- [17] M.V. Bocharinkova, A. S. Saprykin, V.A. Kiktenko, A. S. Adamov, "Developing methods to assess damage from the spread of malware at enterprises", IT-Security Conference for New Generation, Moscow, Russia, 28-29 April 2009.
- [18] X. Wang, M. Tehranipoor and J. Plusquellic, "Detecting Malicious Inclusions in Secure Hardware: Challenges and Solutions", International Workshop on Hardware Oriented Security and Trust, 2008, pp. 15-22.
- [19] I. Verbaudhede, P. Schaumont, "Design methods for Security and Trust", DATE'07, 2007.
- [20] F. Wolff, C. Papachristou, S. Bhunia, R. S. Chakraborty, "Towards Trojan-Free Trusted ICs: Problem Analysis and Detection Scheme", DATE'08, 2008.

ANALYSIS OF DATA MINING VISUALIZATION TECHNIQUES USING ICA AND SOM CONCEPTS

K.S.RATHNAMALA,¹ Dr.R.S.D. WAHIDA BANU²

¹ Research Scholar of Mother Teresa Women's University, Kodaikanal

² Professor & Head, Dept. of Electronics & Communication Engg., GCE.

This research paper is about data mining (DM) and visualization methods using independent component analysis and self organizing map for gaining insight into multidimensional data. A new method is presented for an interactive visualization of cluster structures in a self-organizing Map. By using a contraction model, the regular grid of self-organizing map visualization is smoothly changed toward a presentation that shows better the proximities in the data space. A Novel Visual Data Mining method is proposed for investigating the reliability of estimates resulting from a Stochastic independent component analysis (ICA) algorithm. There are two algorithms presented in this paper that can be used in a general context. Fast ICA for independent binary sources is described. The model resembles the ordinary ICA model but the summation is replaced by the Boolean Operator OR and the multiplication by AND. A heuristic method for estimating the binary mixing matrix is also proposed. Furthermore, the differences on the results when using different objective function in the FastICA estimation algorithm is also discussed.

KEY WORDS:

Independent component analysis, Self organizing map, Vector quantization, patterns,

Agglomerative hierarchical methods, Time series segmentation, Finding patterns by proximity, Clustering validity indices, Feature selection and weighing Fast ICA.

1. INTRODUCTION

The tasks that are encountered within data mining research are predictive modeling, descriptive modeling, discovering rules and patterns, exploratory data analysis, and retrieval by content. Predictive modeling includes many typical tasks of machine learning such as classification and regression. Descriptive modeling that is ultimately about modeling all of the data e.g., estimating its probability distribution. Finding a clustering, segmentation or informative linear representation are common subtasks of descriptive modeling. Particular methods for discovering rules and patterns emphasize finding interesting local characteristics and patterns instead of global models.

Descriptive data mining techniques for data description can be divided roughly into three groups:

Proximity preserving projections for (visual) investigation of the structure of the data.

Partitioning the data by clustering and segmentation .

Linear projections for finding interesting linear combinations of the original variables using principal component analysis and independent component analysis.

A clustering is a partition of the set of all data items $C = \{1, 2, \dots, N\}$ into K disjoint clusters

$$C = \bigcup_i^K C_i = \bigcup_i^K 1C_i$$

2.SELF- ORGANIZING MAP

The basic Self-organizing map is formed of K map units organized on a regular $k \times l$ low-dimensional grid-usually 2D for visualization.

Associated to each map unit i , there is a

1. Neighborhood kernel $h(d_{ij}, \sigma(t))$ where the distance d_{ij} is measured from map unit i to others along the grid (output space), and
2. a codebook vector c_i that quantize the data space (input space).

The magnitude of the neighborhood kernel decreases monotonically with the distance d_{ij} . A typical choice is the Gaussian kernel .

Batch algorithm

One possibility to implement a batch SOM algorithm is to add an extra step to the batch K-means procedure.

$$c_i := \frac{\sum_j^K 1 |C_j| h(d_{ij}, \alpha(t)) c_j}{\sum_j^K 1 |C_j| h(d_{ij}, \sigma(t))}, \forall i$$

A relatively large neighborhood radius in the

beginning gives a global ordering for the map. The kernel width $\sigma(t)$ is then decreased monotonically along with iteration steps which increases the flexibility of the map to provide lower quantization error in the end. If the radius is run to zero, the batch SOM becomes identical to K-means.

The batch SOM is a computational short-cut version of the basic. Despite the intuitive clarity and elegance of the basic SOM, its mathematical analysis has turned out to be rather complex. This comes from the fact that there exists no cost function that the basic SOM would minimize for a probability distribution .

In general, the number of map codebook vectors governs the computational complexity of one iteration step of the SOM. If the size of the SOM is scaled linearly with the number of data vectors, the load scales to $O(MN^2)$. But on the other hand, the selection of K can be made following, e.g., $\sqrt{N}$ as suggested in and the load decreases to $O(MN^{1.5})$. It is suggested that the SOM Toolbox applies to small to medium data sets up to, say, 10 000-100 000 records. A specific problem is that the memory consumption in the SOM Toolbox grows quadratically along with the map size K .

In practice, the SOM and its variants have been successful in a considerable number of application fields and individual applications. In the context of this paper interesting application areas close to VDM include

Visualization and UI techniques especially in information retrieval, and exploratory data analysis in general.

Context-aware computing.

Industrial applications for process monitoring and analysis.

Visualization capabilities, data and noise reduction by topologically restricted vector quantization and practical robustness of the SOM are of benefit to data mining. There are also methods for additional speed-ups in the SOM for especially large datasets in data mining and in document retrieval applications.

The SOM framework is not restricted to Euclidean space or real vectors. A variant of the SOM in a non-Euclidean space or real vectors. A variant of the SOM in a non-Euclidean space is presented to enhance modeling and visualizations of hierarchically distributed data. This method uses a fisheye distortion in the visualization. Also self-organizing maps and similar structures for symbolic data exist and have been applied also to context-aware computation.

3. AGGLOMERATIVE HIERARCHICAL METHODS:

Some clustering methods construct a model of the input data space that inherently would allow classifying a new sample into some of the determined clusters. K-means partition the input data space in this manner. Some other methods

merely provide a partition of the items in the sample: the agglomerative hierarchical methods provide an example of this case.

The family of partitional methods is often opposed to the hierarchical methods. Agglomerative hierarchical methods do not aim at minimizing a global criteria for partitioning, but join data items in bigger clusters in a bottom-up manner. In the beginning, all samples are considered to form their own cluster. After this, at N-1 steps the pair of clusters having minimal pairwise dissimilarity δ are joined, which reduces the number of remaining clusters by one. The merging is repeated until all data is in one cluster. This gives a set of nested partitions and a tree presentation is quite a natural way of representing the result.

Here we list the between-cluster dissimilarities δ of some of the most common agglomeration strategies the single linkage (SL), complete linkage (CL) and average linkage (AL) criteria.

$$\delta_1 = \delta_{SL} = \min_{i \in C_k, j \in C_1} d_{ij}$$

$$\delta_2 = \delta_{CL} = \max_{i \in C_k, j \in C_1} d_{ij}$$

$$\delta_3 = \delta_{AL} = \frac{1}{C_k C_1} \sum_{i \in C_k} \sum_{j \in C_1} d_{ij}$$

Where $C_k, C_1, (k \neq 1)$ are any two distinct clusters. SL and CL are invariant for monotone transformations of dissimilarity. SL is reported to be noise sensitive but capable of producing elongated or chained clusters while CL and AL

tend to produce more spherical clusters. If similarities are used instead, the merging occurs for maximum pairwise cluster similarity.

4. TIME SERIES SEGMENTATION:

In addition to the basic cluster analysis tasks, other clustering methods that include auxiliary constraints are also discussed here. The time series segmentation where the data items have some natural order, e.g., time, which must be taken into account; a segment always consists of a sequence of subsequent samples of the time series.

A K-segmentation divides X into K segments C_i with K-1 segment borders $C_1, \dots, C_{K-1}$ so that

$$C_1 = [x(1), x(2), \dots, x(c_1)], \dots, C_K = [x(c_{K-1}+1), x(c_{K-1}+2), \dots, x(c_N)] \text{ Eq -1}$$

This is the basic time series segmentation task where each segment is considered to emerge from a different model; Furthermore, we consider the case where the data to be segmented is readily available.

As in the basic clustering task, we wish to minimize some adequate cost function by selection of the segment border. We stay with costs which are sums of individual segment costs that are not affected by changes in other segments. An example of such a function is an SSE cost function like that of Eq-1 where c_i is the mean vector of data vectors in segment C_i . There is, of course, a fundamental difference between time series

segmentation with SSE cost and vector quantization. In vector quantization, the borders of the nearest neighbors regions V_i are defined by the codebook vectors, whereas in segmentation, the mean vectors, C_i are determined by the segments C_i but cannot directly be used to infer the segment borders.

Minimizing the cost in Eq- 1 for segmentation aims at describing each segment by its mean value. It may also be seen as splitting the sequence so that the (biased) sample variance computed by pooling the sample variances of the segments together is minimal.

Algorithms

The basic segmentation problem can be solved optimally using dynamic programming. The dynamic programming algorithm finds also optimal 1,2,... K-1 segmentations while searching for an optimal K-segmentation. The computational complexity of dynamic programming is of order $O(KN^2)$ if the cost of a segmentation can be calculated in linear time. It may be too much when there are large amounts of data.

Another class are the merge-split algorithms of which the local and global iterative replacement algorithms (LIR and GIR) resemble the batch K-means in the sense that at each step they change the descriptors of partition (Segment borders vs. codebook vectors) to match with a necessary condition of local optimum. The LIR gets more easily stuck in bad local minima, and the

GIR was considerably better in this sense, yet still sensitive to the initialization. The GIR and LIR algorithms can be seen as variants of the “Pavlidis algorithm” that changes the borders gradually toward a local optimum.

The test procedures use random initialization for the segments. As in the case of K-means, the initialization matters, and it might be advisable to try an educated guess for initial positions. One possibility to create a more effective segmentation algorithm is to combine several greedy methods. For example, the basic bottom-up and top-down methods can be fine-tuned by merge-split methods.

Applications

Time series and other similar segmentation problems arise in different applications, e.g., in approximating functions by piecewise linear functions. This might be done for the purpose of simplifying or analyzing contour or boundary lines. Another aim, important in information retrieval, is to compress or index voluminous signal data. Other applications in data analysis span from phoneme segmentation into finding sequences in biological or industrial process data.

5. VECTOR QUANTIZATION:

Suggested by intuitive aim of the basic clustering task, adequate global clustering criteria can be obtained by minimizing / maximizing a function of

within-cluster dispersion (Scatter) D_W , between cluster dispersion D_B , and their sum, the total dispersion D_T , that is constant and independent of the clustering. For data in a Euclidean space.

$$D_W = \sum_{i=1}^K D_W(i), \quad D_W(i) = \sum_{j \in C_i} (x(j) - ci) (x(j) - ci)^T$$

$$D_B = \sum_{i=1}^K / C_i / ci - c) (ci - c)^T$$

$$D_T = D_W + D_B = \sum_{i=1}^N (x(j) - c) (x(j) - c)^T$$

Where K is the number of clusters, C_i is the average of the data in cluster C_i , and c is the average of all data. These quantities can be formulated also for a general dissimilarity matrix.

The dispersion matrices can be used as a basis for different cost functions. Two criteria invariant to (non-singular) linear transformations of data based on the dispersion matrices; maximizing trace $D_W^{-1} D_W$. Minimizing $\det(D_W)$ gives the maximum likelihood solution for a model where all clusters are assumed to have a Gaussian distribution with the same covariance matrix.

The aforementioned criteria may be difficult to optimize. Therefore a scale dependent criteria, minimization of trace (D_W) has become popular, presumably because it can be (Suboptimally) minimized with the fast and computationally light K-means algorithm that is shortly described in more detail.

Minimization of trace (D_W) is the same as minimizing the sum of squared errors (SSE)

between a data vector $x(i)$ and the nearest cluster centroid C_j :

$$SSE = \sum_{i=1}^K \sum_{x(j) \in C_i} x(j) - c_i^2$$

The above Eq. is encountered in vector quantization, a form of clustering that is particularly intended for compressing data. In vector quantization, the cluster centroids appearing in the above Eq. are called codebook vectors. The codebook vectors partition the input space in nearest neighbor regions V_i . A region V_i associated with the nearest cluster centroid by

$$V_i = \{x: \|x - c_i\| \leq \|x - c_1\|; \forall j\}$$

(nearest neighbor condition).

Cluster C_i in the above Eq is now the set of input data points that belong to V_i .

K-means

k-means refers to a family of algorithms that appear often in the context of vector quantization. K-means algorithms are tremendously popular in clustering and often used for exploratory purpose. As a clustering model the vector quantizer has an obvious limitation. The nearest neighbor regions are convex, which limits the shape of clusters that can be separated.

We consider only the batch k-means algorithm; different sequential procedures are explained. The batch K-means algorithm proceeds by applying alternatively in successive steps the centroid and

nearest neighbor conditions that are necessary for optimal vector quantization.

1. Given a codebook of vectors c_i $i=1,2,\dots,K$ associate the data vectors into codebook vectors according to the nearest neighbor condition. Now, each code book vector has a set of data vectors C_i associated to it.
2. Update the codebook vectors to the centroids of sets C_i according to the centroid condition. That is, for all i set $c_i := (1/|C_i|) \sum_{j \in C_i} x_j$.
3. Repeat from step 1 until the codebook vectors c_i do not change any more.

When the iteration stops, a local minimum for the quantity SSE is achieved K-means typically converges very fast. Furthermore, when $K \ll N$, K-means is computationally far less expensive than the hierarchical agglomerative methods, since computing KN distances between codebook vectors and the data vectors suffices.

Well known problems with the K-means procedure are that it converges but to a local minimum and is quite sensitive to initial conditions. A simple initialization is to start the procedure using K randomly picked vectors from the sample. A first aid solution for trying to avoid bad local minima is to repeat K-means a couple of times from different initial conditions. More advanced solutions include using some form of stochastic relaxation among other modifications.

6. CLUSTERING VALIDITY INDICES

The clustering methods in this paper do not directly make a decision of the number of clusters but require it as a parameter. This poses a question which number of clusters fits best to the "natural structure" of the data. The problem is somewhat vaguely defined since the utility of clusters is not explicitly stated with any cost function. An approach to solve this is the "add-on" relative clustering validity criteria. Basically, one clusters first the data with an algorithm with cluster number $K = 2, 3, \dots, K_{\max}$. Then, the index is computed for the partitions, and (local) minima, maxima, or knee of the index plot indicate the adequate choice(s) of K .

Two examples of such indices Davies-Bouldin type indices are among the most popular relative clustering validity criteria:

$$I_{DB} = \frac{1}{K} \sum_{i=1}^K R_i, \quad R_i = \max_j \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)}, \quad \forall j, j \neq i$$

where $\Delta(C_i)$ is some adequate scalar measure for within-cluster dispersion and $\delta(C_i, C_j)$ for between cluster dispersion. A simplified variant of this, the R-index (I_R) is

$$I_R = \frac{i}{K} \sum_{k=1}^K \frac{S_k^{in}}{S_k^{ex}}, \text{ where}$$

$$S_k^{in} = \frac{1}{|C_k|^2} \sum_{i,j \in C_k} d_{ij}, \text{ and } S_k^{ex} = \min_l \frac{1}{|C_k||C_l|} \sum_{i \in C_k} \sum_{j \in C_l} d_{i,j} \quad (l \neq k).$$

In preliminary experiments, the R-index gave reasonable suggestions for a sensible number of clusters with a given benchmarking data set.

$$V_{AB} = \min_{k, k \neq l} \left\{ \frac{\delta_A(C_k, C_l)}{\max_m \Delta_B(C_m)} \right\}$$

where δ_A is some between-cluster dissimilarity measure and Δ_B is some measure of within-cluster dispersion (diameter), e.g.,

$$\Delta_1(C_k) = \max_{i, j \in C_k} d_{ij}$$

$$\Delta_2(C_k) = \frac{1}{|C_k|^2 - |C_k|} \sum_{i, j \in C_k} d_{ij}.$$

There are literally dozens of relative cluster validity indices and as is obvious, the selection of the R-index is hardly optimal but a working solution and it is only meant to roughly guide the exploration.

6.1. Finding interesting linear projections

Finding patterns in data can be assisted by searching an informative recoding of the original variables by a linear transformation. The linearity is at the same time the power and the weakness of these methods. On one hand, a linear model is limited, but on the other hand, potentially both computationally more tractable and intuitively more understandable than a non-linear method.

6.2. Independent component analysis

In the basic, linear and noise-free, ICA model, we have M latent variables s_i , i.e., the unknown independent components (or source signals) that are mixed linearly to form M observed signals, variables x_i . When X is the observed data, the model becomes

$$X=AS \quad \text{Eq-2}$$

where A is an unknown constant matrix, called the mixing matrix, and S contains the unknown independent components;

$S = [s(1) \ s(2) \dots s(N)]$ consisting of vectors $s(i)$, $s = [s_1 \ s_2 \dots s_M]^T$. The task is to estimate the mixing matrix A (and the realizations of the independent components s_i) using the observed data X alone. The independent components must have non-Gaussian distributions. However, what is often estimated in practice, is the demixing matrix W for $S = WX$, where W is a (pseudo)inverse of A .

This kind of problem setting is pronounced in blind signal separation (BSS) problems, such as the "cocktail party problem" where one has to resolve the utterance of many nearby speakers in the Same room. Several algorithms for performing ICA have been proposed, and the FastICA algorithm is briefly described in the next section.

6.3. FastICA

The FastICA algorithm is based on finding projections that maximize non-Gaussianity measured by an objective function. A necessary condition for independence is uncorrelatedness, and a way of making the basic ICA problem somewhat easier is to whiten the original signals X . Thereafter, it suffices to rotate the whitened data Z suitably, i.e., to find an orthogonal demixing matrix that produces the estimates for the independent components $S = W*Z$. When the

whitening is performed the demixing matrix for the original, centered data is $W = W* \Lambda^{-1/2} E^T$.

Here, we present the symmetrical version of the FastICA algorithm where all independent components are estimated simultaneously:

1. Whiten the data. For simplicity, we denote here the whitened data vectors by x and the mixing matrix for whitened data with W .

2. Initialize the demixing matrix $W = [w_1^T \ w_2^T \ \dots \ w_M^T]^T$ e.g., randomly.

3. Compute new basis vectors using update rule

$$w_j := E(g(w_j^T x)x) - E(g'(w_j^T x))w_j$$

where g is a non-linearity derived from the objective function J ; in case of kurtosis it becomes $g(u) = u^3$, and in case of skewness $g(u) = u^2$. Use sample estimates for expectations.

4. Orthogonalize the new W , e.g., by $W := W(W^T W)^{-1/2}$.

5. Repeat from step 3 until convergence.

There is also a deflatory version of the FastICA algorithm that finds the independent components one by one. It searches for a new component by using the fixed point iteration (in step 3 of the procedure above) in the remaining subspace that is orthogonal to previously found estimates.

Both practical and theoretical reasons make the FastICA an appealing algorithm. It has very competitive computational and convergence properties. Furthermore, FastICA is not restricted to resolve either super or sub-Gaussian sources of the original

sources as it is the case with many algorithms. However, the FastICA algorithm faces the same problems related to suboptimal local minima and random initialization which appear in many other algorithms-including K-means and GIR. Consequently, a special tool Icaso for VDM style assessment of the results was developed in the course of this paper .

6.4. ICA and binary mixture of binary signals

Next, we consider a very specific non-linear mixture of latent variables, the problem of the Boolean mixture of latent binary signals and possibly binary noise. The mixing matrix $\mathbf{A}^B$, the observed data vectors $\mathbf{x}^B$ and the independent, latent source vectors $\mathbf{s}^B$ all consist now of binary vectors $\in \{0,1\}^M$. The basic model in Eq.-2 is replaced by a Boolean expression

$$x_i^B = \bigvee_{j=1}^n a_{ij}^B \wedge s_j^B, \quad i=1,2\dots M$$

where $\wedge$ is Boolean AND and $\vee$ Boolean OR. Instead of using Boolean operators it could be written $\mathbf{x}^B = U(\mathbf{A}^B \mathbf{s}^B)$ using a step function U as a post-mixture non-linearity. The mixture can be further corrupted by binary noise: exclusive-OR type of noise.

On one hand, the basic ICA cannot solve the problem in the above eqn. The methods for post-non-linear mixtures that assume invertible non-linearity cannot be directly applied either. On the other hand, it seems possible that the basic ICA

could work for data emerging from sources and basis vectors that are "sparse enough". Consequently, we experimented how far the performance of the basic ICA can be pushed, using reasonable heuristics, without elaborating something completely new. In this paper, the experiment can be seen as a feasibility study for using ICA where the data was close to binary. Furthermore, there are similar problems in other application fields, prominently in text document analysis where such data is encountered. Since the basic ICA model is not the optimal choice for handling such problems in general, probabilistic models and algorithms have recently been developed for this purpose.

First the estimated linear mixing matrix $\mathbf{A}$ is normalized by dividing each column with the element whose magnitude is largest in that column. Second, the elements below and equal to 0.5 are rounded to zero and those above 0.5 to one:

$$\hat{\mathbf{A}}^B = U(\hat{\mathbf{A}}\Lambda - T)$$

Where the diagonal scaling matrix Λ has elements

$$\lambda_i = \frac{1}{s \max(\hat{a}_i)} \text{ where}$$

$$s \max(\hat{a}_i) = \begin{cases} \min \hat{a}_i & \text{if } \lfloor \min \hat{a}_i \rfloor > \lfloor \max \hat{a}_i \rfloor \\ \max \hat{a}_i & \text{otherwise.} \end{cases}$$

Where $\max \hat{a}_i$ means taking the maximum and $\min \hat{a}_i$ the minimum element of the column vector

$\hat{a}_i$, Matrix T contains thresholds, here we set $t_{ij} =$

0.5, $\forall i, j$. As supposed, this trick works quite well with sparse data and skewness $E(y^3)$ works better than kurtosis as a basis for the objective function on a wide range of sparsity data, except for noisy data.

Conclusion:

In a nutshell, new ways have been presented to develop data mining techniques using SOM and ICA as data visualization methods e.g., to be used in process analysis, an exploratory method of investigating the stability of ICA estimates, enhancements and modifications of algorithms such as the fast fixed-point algorithm for time series segmentation and a heuristic solution to the problem of finding a binary mixing matrix and independent binary sources. Both time-series segmentation and PCA revealed meaningful contexts from the features in a visual data exploration.

REFERENCES:

1. Alhoniemi, E. (2000). Analysis of Pulping Data Using the Self-Organizing Map. *Tappi Journal*, 83(7):66.
2. Cheung, Y.-M. (2003). k^* -Means: A New Generalized k-Means Clustering Algorithm. *Pattern Recognition Letters*, 24(15):2883-2898.
3. Grabmeier, J. and Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery*, 6(4):303-360.
4. Hoffman, P.E. and Grinstein, G.G. (2002). A Survey of Visualizations for High-Dimensional Data Mining. In Fayyad et al. (2002), chapter 2, pages 47-82.
5. Hyvarinen, A., Karhunen, J., and Oja, E. (20010). *Independent Component Analysis*. Wiley Interscience.
6. Lampinen, J. and Kostiaainen, T. (20020). *Generative Probability Density Model in the Self-Organizing Map*. In Seiffert and Jain (2002), chapter 4, pages 75-92.
7. Ultsch, A. (20030). *Maps for the Visualization of High-Dimensional Data Spaces*. In WSOM2003 (2003). CD-ROM.
8. WSOM2003 (2003). *Proceedings of the Workshop on Self-organizing Maps (WSOM2003)*, Hibino, Kitakyushu, Japan.
9. Grinstein, G.G. and Ward, M.O. (2002). *Introduction to Data Visualization*. In Fayyad et al. (2002), chapter 1, pages 21-45.
10. Kohonen, T. (2001). *Self-organizing Maps*. Springer, 3rd edition.
11. Keim, D.A. and Kriegel, H.-P. (1996). *Visualization Techniques for Mining Large Database: A comparison*. *IEEE Transactions of Knowledge and Data Engineering*.
12. Vesanto, J. (2002). *Data Exploration Process Based on the Self-Organizing Map*.
13. WSO2003 (20030). *Proceedings of the Workshop on Self-Organizing Maps (WSOM2003)*, Hibino, Kitakyushu, Japan.
14. Yin, H. (2001) *Visualization Induced SOM (ViSOM)*. In Allinson, N., Yin, H., Allinson, L., and Slack, J., editors, *Advances in Self-Organizing Maps*.

MOBILE AGENT COMPUTING

MRIGANK RAJYA

Software Engineer

HCL Technologies Ltd.

GURGAON, INDIA

mrigankrajya@gmail.com

ABSTARCT

In a broad sense, an *agent* is any program that acts on behalf of a (human) user. A *mobile agent* then is a program which represents a user in a computer network, and is capable of migrating *autonomously* from node to node, to perform some computation on behalf of the user. In computer science, a mobile agent is a composition of computer software and data which is able to migrate (move) from one computer to another autonomously and continue its execution on the destination computer. Mobile Agent, namely, is a type of software agent, with the feature of *autonomy*, *social ability*, *learning*, and most important, *mobility*. Mobile agent is an agent that migrates from machine to machine in a heterogeneous network at times of its own choosing. An agent is “an independent software program which runs on behalf of a network user”. A mobile agent is a program which, once it is launched by a user, can travel from node to node autonomously, and can continue to function even if the user is disconnected from the network. Examples can be Personal assistant (mail filter, scheduling), Information agent (tactical picture agent), E-commerce agent (stock trader, bidder) and Recommendation agent (Firefly, Amazon.com).

Keywords: *Mobile Agent (M.A), agent, paradigm, life cycle.*

1. WORKING OF MOBILE AGENT

A mobile agent consists of the program code and the program execution state (the current values of

variables, next instruction to be executed, etc.). Initially a mobile agent resides on a computer called the home machine [1-2]. The agent is then dispatched to execute on a remote computer called a mobile agent host (a mobile agent host is also called mobile agent platform or mobile agent server). When a mobile agent is dispatched the entire code of the mobile agent and the execution state of the mobile agent is transferred to the host. The host provides a suitable execution environment for the mobile agent to execute. The mobile agent uses resources (CPU, memory, etc.) of the host to perform its task. After completing its task on the host, the mobile agent migrates to another computer. Since the state information is also transferred to the host, Mobile

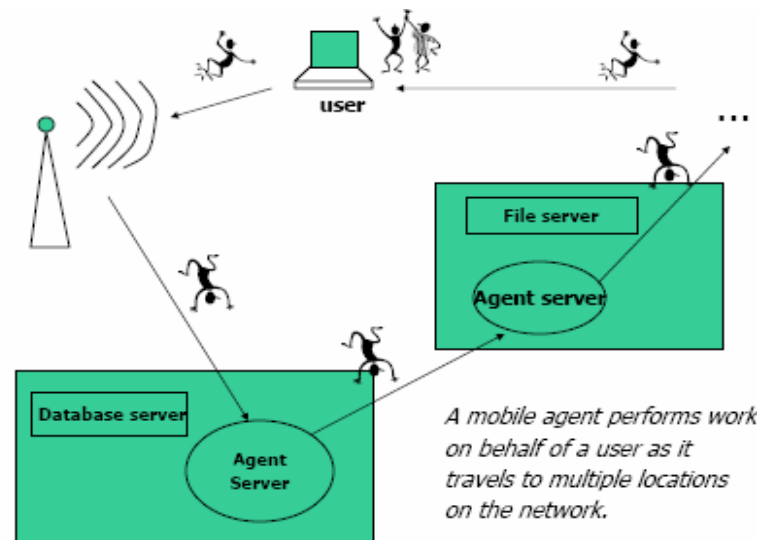


Figure 1

agents can resume the execution of the code from where they left off in the previous host instead of having to restart execution from the beginning. This continues until the mobile agent returns to its home machine after completing execution on the last machine in its itinerary.

2. THE LIFE CYCLE OF MOBILE AGENT

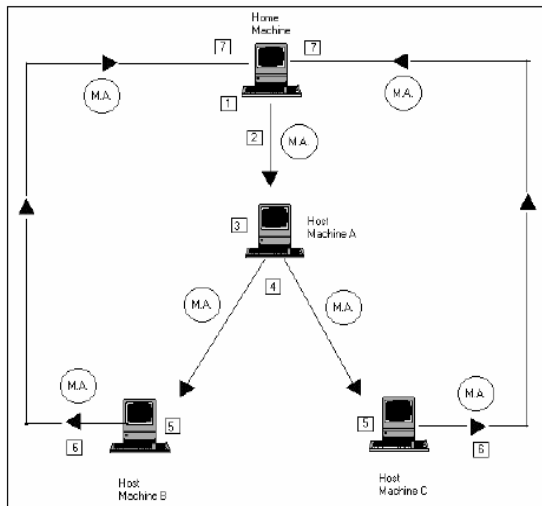


Figure 2

1. The mobile agent is *created* in the Home Machine.
2. The mobile agent is *dispatched* to the Host Machine A for execution.
3. The agent executes on Host Machine A.
4. After execution the agent is *cloned* to create two copies. One (A mobile agent consists of the program code and the program execution state [3-4]. The mobile agent uses resources (CPU, memory etc.) of the host to perform its task) .copy is dispatched to Host Machine B and the other is dispatched to Host Machine C.
5. The cloned copies execute on their respective hosts.
6. After execution, Host Machine B and C send the mobile agent received by them back to the Home Machine.
7. The Home Machine *retracts* the agents and the data brought by the agents is analyzed. The agents are then *disposed*.

3. DESCRIPTION

Part-View of Agent Topology

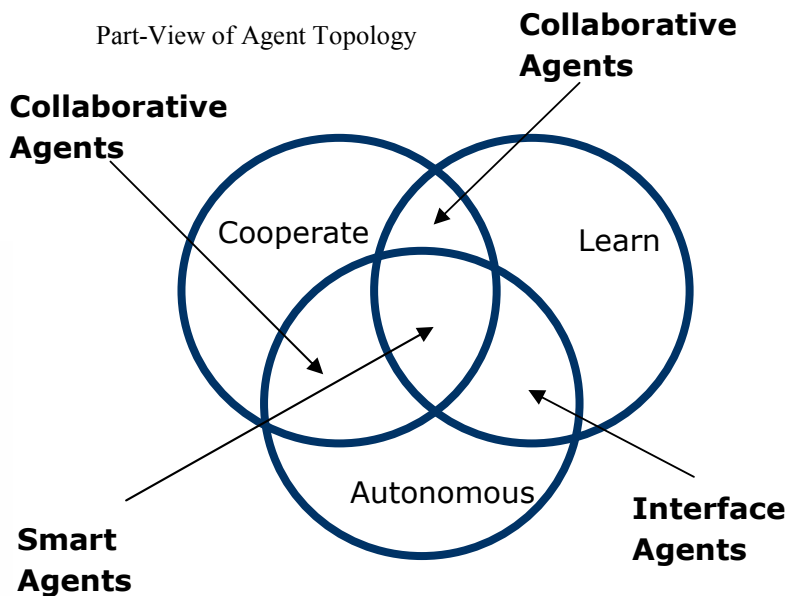


Figure 3

Different types of agents:-

- Agents exist in a multi-dimensional space

- A representative flat-list

1. Collaborative agents
2. Interface agents
3. Mobile agents
4. Information/Internet agents
5. Reactive agents
6. Hybrid agents
7. Smart Agents

- Collaborative Agents

- These emphasize autonomy, and collaboration with other agents to perform their tasks.
- They may need to have “*social*” skills in order to communicate and *negotiate* with other agents.

➤ **Collaborative Agents** example

1. Pleiades Project at CMU.
2. Visitor-Hoster:
 - helps a human secretary to plan the schedule of visitors to CMU
 - matches their interests with the interests and availability of the faculty and staff.
 - organized as a number of agents that retrieve the relevant pieces of information from several different real-world information sources, such as finger, online library search etc.

■ **Interface (Personal) Agents**

- Emphasize autonomy, and learning in order to perform useful tasks for their owners.
- Examples
 1. Personal assistants that handle your appointments
 2. Office Agents in Microsoft Office.
- "Learn" to serve the user better, by observing and imitating the user, through feedback from the user, or by interacting with other agents. The main challenge here is how to assist the user without bothering him, and how to learn effectively.

■ **Information / Internet Agents**

- Focus on
 - helping us to cope with the sheer "*tyranny of information*" in the Internet age.
- Help to
 - manage, manipulate or collate information from many distributed sources.
- Share their
 - respective motivations and challenges

- Functional challenges of managing information.

4. BASIC ARCHITECTURE

An agent server process runs on each participating host. Participating hosts are networked through links that can be low-bandwidth and unreliable. An agent is a serializable object – an object whose data as well as state can be marshaled for transportation over the network. Data marshaling is required for flattening and encoding of data structures, so that they can be sent from one computer to another. An object similarly serialized and transmitted between hosts [5-7]. Upon arrival, the object can be reconstituted and de-serialized, with its execution state restored to when it was serialized, and then the object can resume execution on the newly-arrived host system; i.e. whose execution state can be frozen for transportation and reconstituted upon arrival at a remote site.

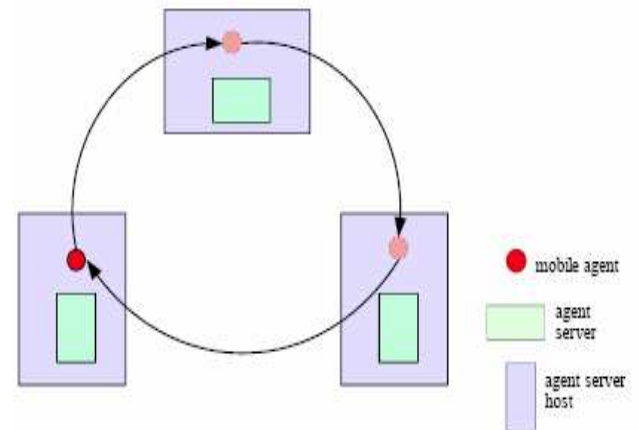


Figure 4

Advantages of the agent paradigm

- Reducing traffic / congestion as agents are smaller in size
- Enhanced security over protected data especially in a broadcast mode

- Avoiding unnecessary data transfer: the transfer of user intention enabling selection of required data and intelligently computed abstraction.
- A modular approach to distributed applications.
- Interoperability: through a new agent layer

5. MOBILE AGENT PARADIGM V/S CLIENT SERVER PARADIGM

In Figure 5 above we have mobile agent paradigm and client server paradigm. The mobile agent paradigm consists of two hosts HOST A and HOST B [8-9]. The client server paradigm consists of client host and server host.

The mobile agent paradigm vs. the client-server paradigm

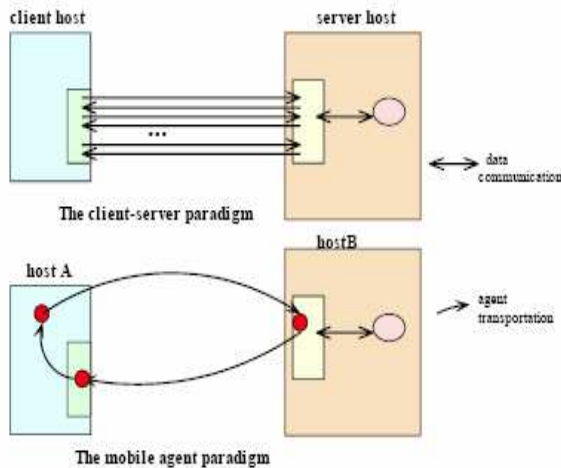


Figure 5

A. Evolution of the “mobile agent” paradigm

The evolution of mobile agent paradigm is explained in Figure 6 as follows.

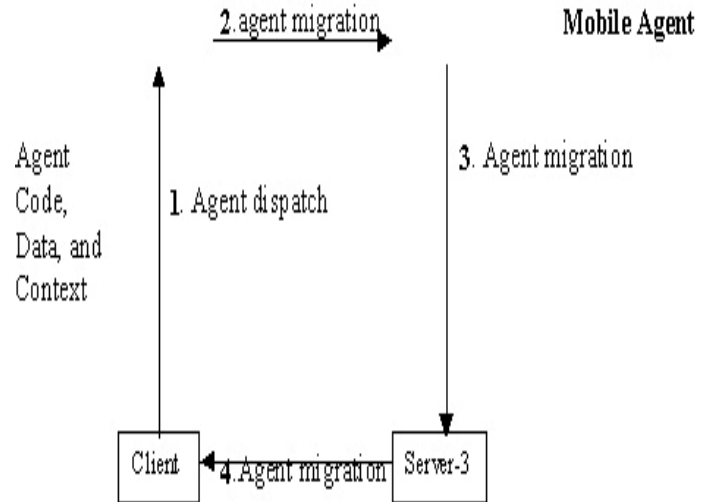


Figure 6

The agent dispatches from the client, then migrates to the server and finally migrates to the client.

B. Properties of mobile agents

Mobile agents have the following unique properties.

1. Adaptive Learning:

Mobile agents can learn from experiences and adapt themselves to the environment. They can monitor traffic in large networks and learn about the trouble spots in the network. Based on the experiences of the agent in the network the agent can choose better routes to reach the next host.

2. Autonomy:

Mobile agents can take some decisions on its own. For example, mobile agents are free to choose the next host and when to migrate to the next host. These decisions are transparent to the user and the decisions are taken in the interest of the user.

3. Mobility:

Mobile agents have the ability to move from one host to another in the network.

6. ATTRIBUTES OF MOBILE AGENT

1. Code
2. State

- a. Execution state
- b. Object state
- 3. Name
 - a. Identifier
 - b. Authority
 - c. Agent system type
- 4. Location

7. ASSUMPTIONS ABOUT COMPUTER SYSTEMS VIOLATED BY MOBILE AGENTS

1. Whenever a program attempts some action, we can easily identify a person to whom that action can be attributed, and it is safe to assume that that person intends the action to be taken.
2. Only persons that are known to the system can execute programs on the system.
3. There is one security domain corresponding to each user; all actions within that domain can be treated the same way.
4. Single-user systems require no security.
5. Essentially all programs are obtained from easily identifiable and generally trusted sources.
6. The users of a given piece of software are restrained by law and custom from various actions against the manufacturer's interests.
7. Significant security threats come from attackers running programs with the intent of accomplishing unauthorized results.
8. Programs cross administrative boundaries only rarely, and only when people intentionally transmit them.
9. A given instance of a program runs entirely on one machine; processes do not cross administrative boundaries at all.
10. A given program runs on only one particular operating system.

11. Computer security is provided by the operating system.

8. THREATS POSED BY MOBILE AGENTS

1) Destruction of

Data, hardware, current environment.

2) Denial of service

- Block execution.
- Take up memory.
- Prevention of access to resources/network.

3) Breach of privacy / theft of resources

- Use of covert channels.
- Obtain/transmit privileged information.

4) Harassment

Display of annoying /offensive information and screen flicker.

5) Repudiation- Ability to deny an event ever happened.

6) Denial of service.

7) Unauthorized use or access of code/data.

8) Unauthorized modification or corruption code/data.

9) Unauthorized access, modification, corruption, or repeat of agent external communication.

9.PROTECTION METHODS AGAINST MALICIOUS MOBILE AGENTS

1. Authenticating credentials

- certificates and digital signatures

2. Access Control and Authorization

- Reference monitor
- security domains

- policies
- 3. Software-based Fault Isolation
 - Java's "sandbox"
- 4. Monitoring
 - auditing of agent's activities
 - setting limits
- 5. Proxy-based approach to host protection
- 6. Code Verification - proof-carrying code

10. MOBILE AGENT SYSTEMS

- Mission oriented single agents: *ex. Information integration in hypermedia or pre-routing congestion awareness.*
- Multiple agent - single agency: *ex. Dynamic routing and network mapping, hidden to end users.*

11. BENEFITS OF MOBILE AGENTS

- Bandwidth conservation
- Reduction of latency
- Reduction of completion time
- Asynchronous(disconnected) communications
- Load balancing
- Dynamic deployment

12. DISADVANTAGES OF MOBILE AGENTS

- The main drawback of mobile agents is the security risk involved in using mobile agents .Security risks in a mobile computing environment are twofold.
- Firstly a malicious mobile agent can damage a host. A virus can be disguised as a mobile agent and distributed in the network causing damage to host machines that execute the agent.

- On the other hand a malicious host can tamper with the functioning of mobile agent.
- To illustrate this scenario consider a mobile agent that visits the servers of several airlines to buy a ticket for the lowest price. A malicious airline server can try to obtain sensitive price information from the mobile agent.
- The malicious server may tamper with the mobile agent and increase the prices quoted by other airlines thereby giving it an unfair advantage.
- Some servers may even try to steal credit card numbers from mobile agent.

13. APPLICATIONS OF MOBILE AGENTS

I. TECHNICAL REPORTS

II. MILITARY

- III. E-COMMERCE - Mobile agents can travel to different trading sites and help to locate the most appropriate deal, negotiate the deal and even finalize business transactions on behalf of their owners. A mobile agent can be programmed to bid in an online auction on behalf of the user. The user himself need not be online during the auction.

- IV. MOBILE COMPUTING - Mobile agents can travel to different trading sites and help to locate the most appropriate deal, negotiate the deal and even finalize business transactions on behalf of their owners [10-12]. A mobile agent can be programmed to bid in an online auction on behalf of the user. The user himself need not be online during the auction.

- V. PARALLEL COMPUTING - Solving a complex problem on a single computer takes a lot of time. To overcome this, mobile agents can be written to solve the problem. These agents migrate to computers on the network, which have the required resources and use them to solve the problem in parallel thereby reducing the time required to solve the problem.

- VI. DATA COLLECTION - Consider a case wherein, data from many clients has to be processed. In the traditional client-server model, all the clients have to send their data to the server for processing resulting in high network traffic. Instead mobile agents can be sent to the individual clients to process data and send back results to the server, thereby reducing the network load.
- VII. INFORMATION RETRIEVAL
- VIII. MONITORING
- IX. SHAREWARE
- X. VIRTUAL MARKET PLACE

14. FUTURE SCOPE

Ad Hoc networks

- Mobile nodes that can be envisioned as routers for message transfer: ex. Laptops
- Each node installed with a transceiver for two way communication in restricted bandwidth: signal strength may be adaptively variable.
- Disconnection among caller & responder viewed as effect of link failure.
- Purely distributive control with no hierarchy at the systems level
- Active nodes transmit periodic beacons for neighborhood information updates.

15. REFERENCES

■ WEBSITE REFERENCES

1. Mysimon.com
2. D'Agents: <http://agent.cs.dartmouth.edu/>
3. Tryllian: <http://www.tryllian.com>
4. Aglets: <http://www.trl.ibm.co.jp/aglets>
5. en.wikipedia.org/wiki/Mobile_agent

6. www.ias.ac.in/resonance/July2002/pdf/July2002p35-43.pdf
7. doi.ieeecomputersociety.org/10.1109/MIS.2006.120
8. www.springerlink.com/index/u2172057t37mp258.pdf
9. netresearch.ics.uci.edu/.../agentos/.../hohl-efficient-code
10. www.cetus-links.org/oo_mobile_agents.html
11. www.objs.com/agent
12. <http://www.cs.dartmouth.edu/~dfk/papers/aslam:position.ps.gz>

■ RESEARCH PAPERS

1. **aslam:position** (BibTeX entry)
Jay Aslam and Marco Cremonini and David Kotz and Daniela Rus. **Using Mobile Agents for Analyzing Intrusion in Computer Networks**. In *Proceedings of the Workshop on Mobile Object Systems at ECOOP 2001*, July, 2001.
2. **cybenko:functional** (BibTeX entry)
George Cybenko and Guofei Jiang. **Matching Conflicts: Functional Validation of Agents**. In *AAAI Workshop of Agent Conflicts*, pages 14-19, Orlando, Florida, July, 1999. AAAI Press.

Map Reduce for DC4.5 and Ensemble Learning in Distributed Data Mining

Dr.E.Chandra

Research Supervisor and Director
Department Of Computer Science
D J Academy for Managerial Excellence
Coimbatore, Tamilnadu, India.
crcspeech@gmail.com

P.Ajitha

Research Scholar and Assistant Professor
Department Of Computer Science
D J Academy for Managerial Excellence
Coimbatore, Tamilnadu, India.
ajitha@y7mail.com

Abstract— MapReduce one of the distributed computing techniques is integrated with decision tree classifier C4.5 in the distributed environment and ensemble learning with its classifier. This paper proposes an algorithm to classify, predict data using MapReduce for DC4.5 with ensemble learning. Proposed algorithm increases the accuracy and scalability of the data. Noise handling in the decision trees with respect to the distributed data is also handled here.

Keywords C4.5, Distributed Decision Trees, MapReduce, Ensemble learning

I. INTRODUCTION

Classification of the decision trees plays a major role in the distributed and centralised environment. Centralised classification techniques are not efficient in handling large volumes of data. Deluge information is available in today's world and it had become necessitated to analyse the information and mine knowledge. For these aspects distributed classification techniques comes in handy. One of the classification techniques C4.5 in distributed environment is discussed here and also MapReduce which comes in handy in the distributed environment.

C4.5 decision tree algorithm is one of the most popular techniques used to predict and classify the data. On massive handling of large data sets may lead to biased selection and chances of missing the value are high. C4.5 selects the unbiased values and prunes the tree in the times of over fitting. Handling discrete and missing values with precision is also possible through C4.5. In a distributed environment, C4.5 selects the attributes across the environment, but the inherent disadvantage of the C4.5 is it is an unstable classification so another learning techniques of ensemble learning and MapReduce is utilised.

In Ensemble learning, base classifiers are constructed from the data sets. New data is classified by combining the predictions of the base classifiers. Ensemble paves way to combining and perturbing many methods to increase the accuracy of and

improves the scalability in a significant way. Ability to Generate multiple versions of the classifiers by the training set or any method or considering any set of parameters.

SPRINT algorithm[3] achieve the scalable performance in the parallel and distributed environment. Prodromidis[4] in the meta learning discusses in respect to the distributed environment. JAM[5] in the grid and cloud computing environment and Weka[6] for grid enabled cross validation and testing. These are the few algorithms already specifies for the parallel and distributed environment with ensemble methods.

Map Reduce [1] proposed by Google for handling massive large data sets. There are two primary functions (1) a Map Function (2) Reduce Function for the distributed, parallel and cloud computing environment. Basic work of map reduce is to iterate the input, compute key – value pairs for each part of the input, group all intermediate values by key, then again iterate over the resulting groups and finally reduce each group. Implementation issues like fault tolerance, load balancing and performance are also be able to handle in the Map Reduce environment. Series of machine learning techniques are used in MapReduce to efficiently solve the problems arises in large scale distributed environment.

Recently ensemble learning had become popular , because of it is inherent nature to train many learnings and combine or group the results. For the distributed environment ensemble learning can increase the accuracy and reduce the computing efforts. Utilising the advantages of Map Reduce, ensemble and C4.5 , a new classification algorithm DEC4.5-MR is proposed here. Distributed ensemble C4.5 with map reduce algorithm will classify , construct and predict the data.

Rest of the paper is organised as follows. Section 2 discuss the related work exists for C4.5, ensemble, Map Reduce in distributed environment. Section 3 examines the distributed C4.5 with ensemble and Map Reduce. Section 4 the experimental evaluation and Section 5 described the conclusions of this paper

II. RELATED WORK

A. C4.5 Method

C4.5 is a successor of CLS and ID3. Generates the classifier in form of decision trees and generate rules based on the results. Divide and conquer strategy is utilised in the C4.5 which is described in the figure 1. Two heuristic evaluation methods like information gain and gini index is used in the C4.5 both the heuristic technique has the ability to handle discrete, missing values and precision of the data handling will be high. Numeric or Nominal data can also be considered. After the construction pruning can be done to avoid over fitting by using post pruning pessimistic method. Suppose for N nodes, E will be the errors that occurs and most frequent node is set and can be experimented by Bernoulli experiments equation specifies the same. Pessimistic error condition can be computed using equation 2.

$$pr \left[\frac{f - q}{\sqrt{q(1-q)N}} > z \right] = c \quad \text{---- (1)}$$

$$e = \frac{f + \frac{z^2}{2N} + Z \sqrt{\frac{f}{N} + \frac{f^2}{N} + \frac{Z^2}{4N^2}}}{1 + \frac{z^2}{N}} \quad \text{----(2)}$$

The algorithm for distributed DC4.5 decision tree is specified in figure1 .

Input: A training sets S , a node T ;
Output: A decision tree with the root T ;
1: If the instances in S belong to the same class or the amount of instances in S is too few, set T as leaf node and label the node T with the most frequent class in S ;
2: Otherwise, choose a test attribute X with two or more outcomes based on a selecting criterion, and label the node T with X ;
3: Partition S into subsets $S_1, S_2, \dots, S_n$ according to the outcomes of attribute X for each instance; generate T 's n children nodes $T_1, T_2, \dots, T_n$;
4: For every group (S_i, T_i) , build recursively a subtree with the root T_i .
5. Each and every group is integrated and global tree is generated

Fig1 Algorithm for DC4.5 decision tree

Algorithm for DC4.5 of the distributed decision tree discusses how to use C4.5 in distributed environment.

B. Ensemble Learning in DDM

Ensemble learning methods in distributed data mining are of 2 methods one is either mining inherently distributed data and scaling up ensemble methods based on partitioning and combining results. When the nature of data is geographically distributed and handling it in centralised environment is not efficient without perturbing the conventional methods. Majority voting and weighted voting both can be combined in distributed decision trees. So ensemble technique allows the partition of vertical partitioning otherwise called as heterogeneous data[7]. This ensemble technique increases accuracy, reduces the processing time in compared with centralised prediction. Scalability issues are also handled in the ensemble methods

Bagging, boosting and sub space are some of the algorithms in ensemble techniques. Here, bagging one of the most and popular classifier is utilised here as it scales well in distributed environments[9]. Bagging is well able to handle[8] noise in the data

Bagging ensemble algorithm

Input: L : a classification method, D : a training data set, m : the number of base classifiers;

Output: $\phi(*)$: an ensemble classifier;

1: for $i = 1$ to m

2: $D' =$ bootstrap sampling from the set D ;

3: $\phi_i(*) = L(D')$;

4: end for;

/* Y is a nominal class label set */

5: $\phi(*) = \arg \max_{y \in Y} \sum_{i=1}^m \phi_i(*) = y$;

Fig 2. The Bagging Ensemble Algorithm.

C. MapReduce Distributed Computing Model

The following architecture specifies the MapReduce distributed computing Model and the operation of the two functions Map and Reduce. The architecture of the MapReduce splits the Map and Value pair and local tasks player with the necessity Jobs Scheduling and work.

Map function can be highly parallelized and so is the Reduce function. Map functions usually deal with parallelized and distributed sub-missions. Reduce function usually collect the sub-missions and parallelize or distributed it according to the need[10]. Computing process of Map Reduce is short(i) Divide into large sets (ii) each (or several) data sets handled in one cluster for intermediate processing (iii) intermediate is combine with final cluster. The Map function read input data sets with the format of $\langle key1, value1 \rangle$. After the analysis, it generates an intermediate result $\langle key2, value2 \rangle$ and submits it to a Reducer; then the Reduce function combines the results to get a final list $\langle key3, value3 \rangle$ according to the list $\langle key2, value2 \rangle$.

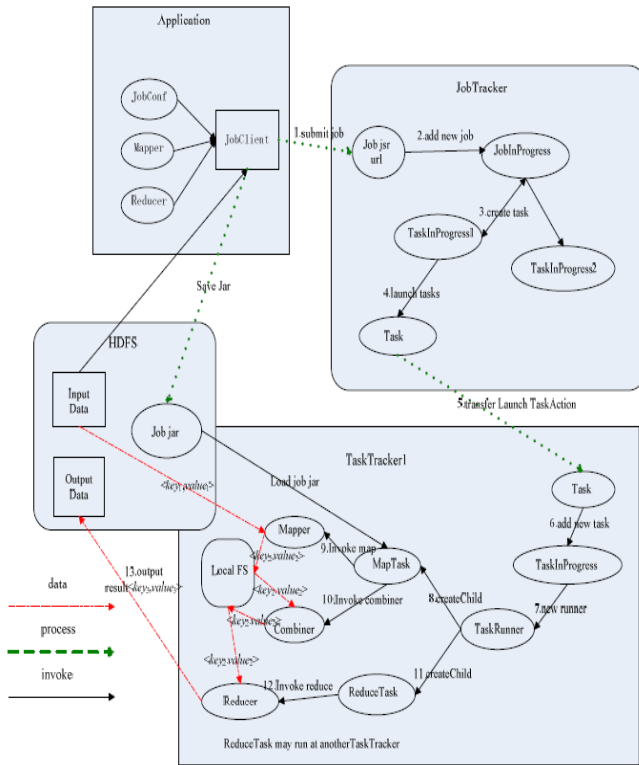


Fig 3-Hadoop MapReduce framework Architecture

During the Map process, in order to improve the combination efficiency, a Combiner can be used which has similar function as Reducer to reduce at local. The transformation between the input and the output looks as follows [6].

$map(key1, value1) \rightarrow (key2, value2) []$
 $reduce(key2, value2 []) \rightarrow (key3, value3) []$

Hadoop is an implementation of the MapReduce parallel computing model of the open source framework for distributed programming. With the help of Hadoop, programmers can easily write parallel and distributed programs. It runs in computing clusters to deal with massive data [11]. The basic components of an application on Hadoop's MapReduce include a Mapper and a Reducer class, as well as a program to create a JobConf. Some applications also include a Combiner class which is actually the implement of the Reducer on local.

Hadoop implements a distributed file system, referred to HDFS. HDFS has the characteristic of high fault-tolerant, and is designed for deployed on low-cost hardware. It provides high throughput to access the data of applications, which is suitable for an application with large data sets. HDFS relax the requirements of POSIX, allowing streaming access to data in the file system. In addition, Hadoop implements the MapReduce distributed computing paradigm. MapReduce splits the mission of an application into small blocks of work. HDFS establishes multiple replicas of data blocks for

reliability and places them on compute nodes on server groups. So MapReduce can handle these data on the associated nodes. Figure 3 shows the Hadoop MapReduce framework [10].

III. PROPOSED ALGORITHM

A. Distributed C4.5 with ensemble and MapReduce

The proposed algorithm DC4.5 with ensemble and Map reduces has 3 phases. Three phases are partition Phase/the map, Build base classifier phase and Reduce/Ensemble phase. Divide data sets D into n subsets of $\{D_1, D_2, \dots, D_n\}$ and users determine the value n . In the first phase of Map phase a Base classifier BC_i needs to be generated into Classifier C_i with the DC4.5 algorithm. In Reduce/Ensemble phase assemble the n base classifiers to generated final classifier using Bagging.

B. Types of keys and values

The types of sets of key and value of MReC4.5 illustrates as follows

key1 : Text
value1 : Instances
key2 : Text
value2 : Iterator of Classifiers
key3 : Text
value3 : Classifier

key1, key2, key3 are all the Text type offered by Hadoop and their values are the file name associated with the input data set D . In the Partition phase, when the data set D is split into m data sets, according to the input format of the C4.5 algorithm each data set is formatted as value1 with the Instances type. In the Map phase, we build a classifier model with the C4.5 algorithm and obtain a classifier model set value2 which belongs to the Iterator of Classifiers type; In the Reducer phase, we assemble classifiers from value2 to obtain a classifier model value3 with the Classifier type.

C. Map/Reduce Phase

Figure 4 specifies the proposed algorithm for the Map operations in respect to the C4.5 algorithm. A change is done in the original proposed algorithm to Map and reduce for the key value pairs

```
function mapper(key, value)
/* Build base-classifier */
1: Build a C4.5 Classifier  $c$  with the data set value;
/* Submit intermediate results */
2: Emit(key, c);
3. Generate  $\{D_1, D_2, \dots, D_n\}$  for all subsets
4. Build and Map with  $C_i$ 
5. Integrate into one cluster the (key, c)
```

Fig 4. The Map Operation

In the mapper functions C4.5 classifier is used so that data set value with the key is taken and paired with the distributed data list. After pairing with the *value_list* emit the key. Generating all the subsets of data sets which further adhere to the key is classified. On classifiers of the all subsets of data Build and Map the $C_1, C_2, \dots, C_n$ with the $D_1, D_2, \dots, D_n$.

D. Reduce/Ensemble Phase

Figure 5 specifies the proposed algorithm for reduce operations with the bagging ensemble which eliminates noise and scales the data well.

```
function reducer(key, value_list)  
/* Get each classifier model */  
1: foreach value in value_list  
2: classifiers[i++] = getClassifier(value);  
/* Perform the bagging ensemble */  
3: c = baggingEnsemble(classifiers);  
4: Emit(key, c);  
5. Integrate  $C_i$  and value_list  
6. Generate the model and predict the results.  
7. Reduce the value_list based on the key pair  
8.  $C_i$  and key with the model .  
9. select only the key which provides the closest combination  
to the value_list generated.
```

Fig 5. The Reduce Operation for DC4.5 with ensemble and Map Reduce

value_list specifies the combination of all intermediate lists and submit the final to Map Reduce Framework.

DC4.5 partitions the data handled and after partitioning the data sets into required and necessities format, the data are transferred to next phase of build /classifier phase. In classifier phase the classification technique of decision trees in distributed data are utilised. Here, the gini index and information gain are made use of. After the classifier phase the map operations are build and mapped to the cluster with the reference to the key value pair.

Next Phase of Reduce/Ensemble phase each of the value are classified and bagging Ensemble technique is used for the further classification and model generation. After generating the models the results are predicated according to the key-value pairs and integrated based on the key.

IV. EXPERIMENTAL EVALUATION

The computational complexity of the algorithm is $O((n+m)\log n)$ where the n and m specifies the values and accuracy. $\log n$ specifies how it scales well compared to the centralized decision trees. Based on the other methods this increases greater accuracy as it takes the $O(n)$ where n is the large data sets size.

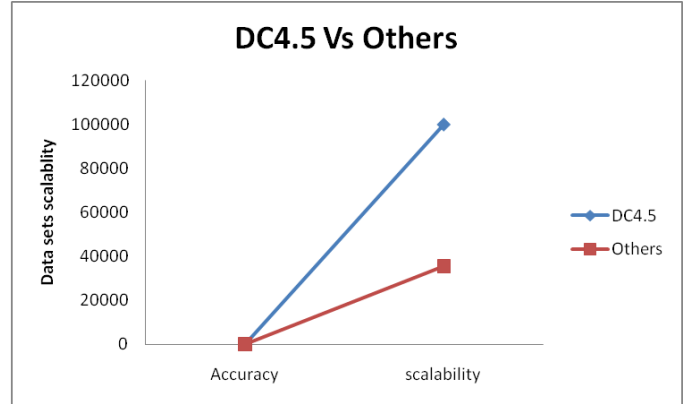


Fig 6 Accuracy Vs Scalability chart for DC4.5 and others

Hadoop MapReduce takes the various data sets with the attributes characteristics of both numeric and categorical attributes. On this experimental evaluation computational complexity of the each MapReduce phase on the basis of the $O(n \log n^2)$ where n specifies the data sets and the accuracy handling based on the emit of key-value pair.

5.CONCLUSIONS

This paper deals with the distributed C4.5 in the ensemble learning for the distributed environment utilizing the Map reduce framework. The proposed algorithm increases accuracy as well as scalability of the data handled. Inherent nature of the number of data handling in the Distributed Decision Trees can be improved based on the *value_list* of the key pairs in the MapReduce environment.

Comparing to the centralized decision trees the algorithm proposed can be utilised efficiently in terms of the processing time and computation cost. Future work may be enhanced well with the detailed implementation of the same on the Hadoop MapReduce framework so that number of millions of data with the characteristics can also be integrated.

REFERENCES

1. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *OSDI'04: Proceedings of the 6th Symposium on Operating System Design and Implementation*. San Francisco, California, USA. USENIX Association, pp. 137-150. December 6-8, 2004.
2. Apache, "Hadoop", <http://lucene.apache.org/hadoop/>, 2006.
3. J. C. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A scalable parallel classifier for data mining," in *VLDB'96: Proceedings of the 22th International Conference on Very Large Data Bases*. Mumbai (Bombay), India, Morgan Kaufmann, pp. 544-555, September 3-6, 1996.
4. A. Prodromidis, P. Chan, and S. Stolfo, "Meta-learning in distributed data mining systems: Issues and approaches," *Advances in Distributed and Parallel Knowledge Discovery*, Vol. 114, 2000.
5. K. Cardona, J. Secretan, M. Georgiopoulos, and G. Anagnostopoulos, "A Grid Based System for Data Mining Using MapReduce", Technical Report, University of Puerto Rico, July 2007.
6. R. Khossainov, X. Zuo, and N. Kushmerick, "Grid-enabled weka: A toolkit for machine learning on the grid," *ERCIM News* No. 59, October 2004.

7. Skillicorn, D.B and S.M.McConnell, 2008. Distributed prediction from vertically partitioned data. *J.Parallel Distributed Computing*, 68:16-36.
8. T. G. Dietterich, "Machine-learning research: Four current directions," *The AI Magazine*, vol. 18, no. 4, 1998, pp. 97-136.
9. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. *Chapman & Hall/CRC*, May 1994.
10. D. Borthakur, The Hadoop Distributed File System: Architecture and Design, *The Apache Software Foundation*, 2007.
11. L. A. Barroso, J. Dean, and U. Holzle, "Web search for a planet: The google cluster architecture", *Micro, IEEE*, vol. 23, no. 2, 2003, pp.22-28.

AUTHORS PROFILE



Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University, Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 15 yrs of experience in teaching including 6 months in the industry. Presently she is working as Director, Department of Computer Applications in D. J.

Academy for Managerial Excellence, Coimbatore. She has published more than 30 research papers in National, International Journals and Conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. Currently 3 M.Phil Scholars and 8 Ph.D Scholars are working under her guidance. She has delivered lectures to various Colleges. She is a Board of studies member of various Institutions. Her research interest lies in the area of Data Mining, Artificial Intelligence, Neural Networks, Speech Recognition Systems, Fuzzy Logic and Machine Learning Techniques. She is an active and



Life member of CSI, Society of Statistics and Computer Applications. Currently she is Management Committee member of CSI Coimbatore Chapter.

P.Ajitha,, received her B.Com from Bharathiar University, Coimbatore in 1998 and received MCA from Bharathiadsan University, Trichy in 2001. She obtained her M.Phil in the area of Data Mining in 2004. She has 9 years of experience in teaching and 3 months of industrial experience. Currently, she is working as Assistant Professor, Department Of Computer Applications, D.J.Academy for Managerial Excellence, Coimbatore and pursuing her Ph.D in Bharathiar University, Coimbatore. She has presented more than 10 research papers in National and International Conferences and published a paper in an International Journal. Her Research Interest lies in Distributed Data Mining, Machine Learning and Artificial Intelligence. She is Life a member of CSI, life member of Institute of Advanced Scientific Research and also a member IASCIT.

A Novel E-Service for E-Government

A. M. Riad, Hazem M. El-Bakry, and Gamal H. El-Adl

Dept. of Information Systems
Faculty of Computer Science and Information Systems, Mansoura University
Mansoura, Egypt

Abstract—Egyptian e-government facilitates and introduces e-services for its partnerships such as citizens, businesses, employees and government itself. The combination of geographic information systems (GIS) as Egyptian new e-service with decision support systems (DSS) is used to help the ministry of finance. Our proposed e-service appears the important geographical criterions that affect the value of tax rates. The committee of tax rates determination for housing units used new e-service in order to reduce the time required to manually check all buildings in the country. So the estimation of the housing tax rates is based on the spatial data from the novel GIS e-service.

Keywords- E-government, E-services, GIS, GPS, Housing Tax Rates.

I. INTRODUCTION

E-government involves using information technology, and especially the Internet, to improve the delivery of government services to citizens, businesses, and other government agencies to interact and receive services from the federal, state or local governments twenty four hours a day, seven days a week. It transforms the public service into electronic service via information and communication technologies (ICT). Nowadays, the rapid development of computer technology as Internet facilitates easy access to data, information, and knowledge sources which are available online. The Egyptian government gives impetus to the development of ICT in Egypt. It concentrates on linking them to the economic and social development of the Egyptian country [1]. This is to move from manual information system to computer based one. The information society should be able to deliver high quality government services to the public where they are and in the format that is suitable for them. Egyptian e-government was created since 2001 for many reasons. First, in order to reach a new level of convenience in government services. Second, to offer citizens the opportunity to share in the decision making process. Third, to greatly improve efficiency and quality of services. Fourth, to enhance the Egyptian economy and help in increasing production. In the last year, the ministry of finance applied a new rules for computing the housing tax rates to increase its ability for payment and to support the other needs [1-2]. The question here is how to determine the tax rates?, How to get it quickly?, and finally how to scan all units in short time?. Here, GIS is used to easy access the all units at the same time and get all the required information about the owners [8-10].

This paper is organized as follows. Section II explains the technology of e-service and reason for using this e-government. The proposed e-service is presented in section III. The combination of GIS with GPS to provide an efficient DSS for determination of housing tax rates is discussed. Finally, conclusions are given in section IV.

II. WHY E-SERVICE?

E-government is now a central theme in information society at all levels: local, national, regional and even global. It can be defined as a transformation of public-sector internal and external relationships through use of ICT to promote greater accountability of the Government, increase efficiency and cost-effectiveness and create a greater constituency participation. Countries of the Asian and Pacific region engage in e-government, as they provide cost-effective government-related information via Web sites and most have already developed a national e-government strategy (often as part of an ICT strategy plan). The emerging economies in the region have already gone one step further in introducing internal information management at various levels of sophistication. However, only a few Governments in the region have successfully implemented a comprehensive set of online public services, and even fewer have backed these operations up with comprehensive knowledge management in ministries and between the various government agencies. Even though, most governments in the region are eager to further benefit from e-government, by improving efficiency and transparency of the public sector, and providing inclusive public services, they may feel that e-government is a concept far removed from their current realities. ICT applications in the public sector can be used as a strategic tool for development and also a response to the current challenges of globalization. For all Governments, e-government was a fundamental complement to the successful implementation of a range of other government policy targets. E-government was clearly linked to the international competitiveness of an economy and was a fundamental driver of economic growth along with monetary, fiscal, labour and trade policies. E-government pushed the limits of traditional government, changing the way in which government functioned and fostering a culture that made the customer and citizen central to everything it did. It involved building an integrated, enabling infrastructure that could meet the requirements of today's environment, while being readily adaptable to new and innovative developments [11].

While the benefits of e-government were growing, there remained a need for a better understanding of the impact and role of e-government. Owing to the tremendous resources required in implementing e-government, the sharing of knowledge and experience could help developing countries in the region to reduce costs and limit unnecessary mistakes. However, there was a need to define an e-government agenda, and give priorities and specific recommendations on how best to move e-government forward. E-government had impacted on all levels of government. Successful economies were those where a central coordinating agency had been formed to oversee the shift to e-government. If there was not a uniform approach, e-government was destined to failure. E-government could have effects on policy and programs objectives through:

- Improved services, e.g. customer satisfaction, burden reduction and savings
- Enhanced economic development
- Improved policy formulation
- Redefined communities
- Increased operational efficiency
- Enhanced citizen participation

Furthermore, e-government could be used as an anchor to drive transformation across the public and private sector and as a tool to drive foreign investment and economic development. It was important not to over emphasize the role of technology – technology was often a large part of cost, and only a small part of success. To ensure success, the following items are needed to be done:

- Become customer-centric
- Learn how to cope with change
- Develop technical infrastructure
- Collaborate for success
- Work across silos, break down traditional, hierarchical structures
- Develop performance measures

All of those elements were necessary for transforming the government. The technological infrastructure was the base upon which other changes could be made. For overall transformation in the government those issues are needed to be examined in the context of one another [11].

The Egyptian ministry of finance introduced new rules for computing the housing tax since 2009. Housing tax rate is determined by some of important criterions. Such criterions are

1. The total numbers of floors.
2. The area of the house.
3. The space around the house.
4. The location of the house.
5. The strategic position which means whether the house is nearing from a global location such as river, garden or public streets/places.

The ministry of finance assigned this complex task to the committee of evaluation and control. Such committee is responsible for calculating the housing tax rates. The duties of this committee are:

1. Collecting data entered by owners
2. Reviewing such data.
3. Checking the correctness of this data.
4. Scanning all buildings, streets and every other location.
5. Investigate the total income for each owner.
6. Computing the ratio between the total income for the owner and the number of members in his family.
7. Deciding about the way of payment along the year.

It is obvious that it will be very difficult to calculate the tax rates. Moreover, it is expected that those procedures will consume long time. The process of calculating the tax is sharable with the help of the owners of units. The owners introduce a real estate tax declaration either electronically on line or by handling it in the offices of Egyptian ministry of finance. The committee can ensure the correctness of the entered data by owner via our novel e-service. The evaluation of any e-service is customer based. Therefore, the determination of the quality of e-services should be performed. This can be done by the customers themselves.

III. The Proposed E-Service

While much of e-government relied on telecommunication innovations such as bandwidth and speed, there was also a need to focus on how to distribute e-government applications to potential users. E-government access was about providing services to citizens and business in ways that they chose to apply to them, at a time appropriate to them. Further, universal access was essential. Therefore, providers must choose the most appropriate delivery channels.

DSS are mechanisms that can be used to provide managers with information needed to make managerial decisions [12-73]. Decision support systems are gaining recognition in the public sector, which seeks solutions to various problems in a number of diverse areas. Many solutions are closely tied to individual fields, such as medicine [74], ecology [75] and spatial planning [76]. Others, in a more general way, are directed towards support in strategic planning and solving problems in management [77-78]. Lately, due to the redirection of politics away from ascertaining public opinion about the functioning of the public sector towards public engagement and cooperation in decision making processes, the number of solutions in the area of e-democracy is increasing [79-82]. Support systems and cooperation in decision making are, however, still used mainly in narrow professional circles and have not found their way to political decision makers or to the public [83]. The challenge of successful implementation of a decision support system in the public sector, with engagement over the whole spectrum of decision making, is still unmet [84-95].

GIS is a specialized information system having all the basic possibilities of an information system as query, reporting as well as data storage and retrieval [76]. A data model is a representation of some real world phenomenon for which information will be stored in a database. Storing information

IV. CONCLUSION

in a database has many benefits such as allowing the user to perform complicated analytical functions and queries, handling large amounts of data, imposing certain rules on the stored data. GIS makes use of attribute data associated with geographical data (spatial data) [4]. Geographical data may be represented as points, lines or polygons. Attribute data can be handled easily using a conventional database management system (DBMS) [3-7]. GIS has the ability to query this spatial data. GIS is defined by its ability to cater for spatial queries. GIS allows you to query and find geographical features using addresses. Moreover GIS is spatial analysis tools [12-73]. GIS is used in our proposed e-service as a tool to help the committee of evaluation and control in determining the values of housing rate taxes perfectly. After many interviews with managers in the Egyptian ministry of finance, they told us about the major important criterions to evaluate the values of tax rates. Table I shows the most important criterions that affect the values of tax rates and its corresponding GIS data in the representation layer.

The novel GIS e-service is used as a tool to support the committee of evaluation and control in order to reduce the time consumed to manually checking every flat in all cities. Fig. 1 describes the block diagram for the steps of our novel GIS e-service. The proposed e-service not only helps the citizen but also all the partners of the Egyptian e-government. The satellite moons will capture all towns in Egypt. Another way is to get the maps from Google Earth. The process of digitizing converts the master map into vector map by determining the important real world criterions as shown in Table I. Examples for these criterions are gardens layers, and rivers. The owners of the housing units access the online tax declaration form and enter the data properties of their units. The committee of evaluation and control can determine the values of the housing tax rates depending on the resulted spatial data. The flow chart of the housing tax computation rates is shown in Fig. 2. Fig. 3 shows the interface of the proposed GIS e-service. The selected location in any city and its different layers are appeared. Fig. 4 clarifies the final segmentation of the real world geographical criterions that affect the values of the tax rates. Spatial query interface for supporting the committee is presented in Fig. 5. Fig. 6 describes the main responsibility of the unit owners which is to enter data of their units like location, size, type (such as apartment, villa or building and so on). The committee of evaluation and control can use latitude and longitude query through our novel GIS e-service to get information about any location or unit. The GPS can be used as manual equipment in order to check the correctness of owner data that is entered into the main property form. GPS finds the actual real latitude and longitude degrees. Fig. 7 presents the results of committee's query about the latitude and longitude of random checking building.

Table II shows the difference between traditional manual service and our novel e-service. The traditional manual service needs more effort than our new e-service. This is because the effort to manually scanning all units is very huge. The novel GIS e-service reduces the time consumed compared with manual checking at every flat in all cities.

A new technique for computing housing tax rates has been presented. It has been shown that such technique facilitates this governmental service for both citizens and the committee of evaluation and control. The values of the tax rates have been estimated in real-time. Furthermore, all of the housing units have been scanned simultaneously. This has been achieved by applying GIS in e-government systems. In addition, it has been proven that the combination of GIS and GPS for DSS has developed the e-services in the Egyptian ministry of finance. Moreover, the tax rate of any flat has been computed accurately according to its location by using the proposed e-service. Compared to the manual computing system for housing tax rates, the required time has been reduced by using our proposed technique for any housing unit in the city. The presented approach can be applied for computing any other types of taxes that depend on the geographical location.

REFERENCES

- [1] Ministry of Communication and Information Technology, "EISI-Government the Egyptian Information Society Initiative for Government Services Delivery," online available at http://www.mcit.gov.eg/Egy_vis_infoc 3 1.asp
- [2] Azab, N. A., Kamel, S. and Dafoulas, G. "A suggested Framework for Assessing Electronic Government Readiness in Egypt," *Electronic Journal of e-Government*, Vol.7, No.1, pp11-28, 2009.
- [3] A.Goings, D.Young, S. H. Hendry "Critical Factors in the Delivery of E-government Services," *Conference of the International information Manager Association*, Vol.3, No.3, 2004.
- [4] "GIS for managing survey data "online available at http://www.fig.net/pub/jakarta/papers/ts_20/ts_20_3_majeed_parker.pdf
- [5] Backus, M. ,"E-Governance and Developing Countries, Introduction and examples," *Research Report*, No.3, 2001
- [6] Sharma, S. K. and Gupta, J. N. D. "Building Blocks of an E-Government – A Framework," *Journal of Electronic Commerce in Organization*, Vol.1, No.4, pp 34-48, 2003.
- [7] Peter Salhofer, D. Ferbas, "A pragmatic Approach to the Introduction of E-Government," *Proc.8, International Government Research Conference*, 2007.
- [8] Zhiyuan Fang, "E-government in Digital Era: Concept, Practice, and Development," *International Journal of the Computer, the Internet and Management*," Vol.10, No.2, pp 1-22, 2002.
- [9] Shivakumar Kolachalam, "An Overview of E-government," *International Symposium on learning Management and Technology Development in the Information and Internet Age*. Online available at www.ea2000.com, 2003.
- [10] Shailendra C. Jain Palvia, Sushil S.Sharma " E.Government and E-Governance: Definitions/Domain Framework and Status around the World," *International Conference on E-governance*, 2007
- [11] Goodchild, M., 1998, *Geographical information systems and disaggregate transportation modeling*, *Geographical Systems*, volume 5, 1998.
- [12] Keen, P. and Scott-Morton, M. "Decision Support Systems: an organizational perspective", Addison-Wesley Publishing 1978.
- [13] Abdelkader ADLA "A Cooperative Intelligent Decision Support System for Contingency Management", *Journal of Computer Science* Vol.2, No.10, pp 758-764, 2006.
- [14] Roger L. Hayen, "Investigating decision support system frameworks", *Issues in Information Systems Journal*, Vol. 2, No.2, 2006.

- [15] Backus, M., "E-Governance and Developing Countries, Introduction and examples", Research Report, No.3, 2001
- [16] Sharma, S. K. and Gupta, J. N. D., "Building Blocks of an E-Government – A Framework", Journal of Electronic Commerce in Organization, Vol.1, No.4, pp 34-48, 2003.
- [17] P.Salhofer, and D.Ferbas, "A pragmatic Approach to the Introduction of E-Government", Proc.8, International Government Research Conference, 2007.
- [18] Zhiyuan Fang, "E-government in Digital Era: Concept, Practice, and Development", International Journal of the Computer, the Internet and Management," Vol.10, No.2, pp 1-22, 2002.
- [19] Shivakumar Kolachalam, "An Overview of E-government", International Symposium on learning Management and Technology Development in the Information and Internet Age, online available at www.ea2000.com, 2003.
- [20] Goodchild, Michael F., (2010). Twenty years of progress: GIS Science in 2010. JOURNAL OF SPATIAL INFORMATION SCIENCE Number 1 pp. 3–20 doi:10.5311/JOSIS.2010.1.2. July 27, 2010.
- [21] Chang, K. T. (2008). Introduction to Geographical Information Systems. New York: McGraw Hill. p. 184.
- [22] Wilson, J.P.; Fotheringham, A.S. (2007), The Handbook of Geographic Information Science, Malden, MA: Blackwell Publishing, pp. 3–12.
- [23] Longley, Paul A.; Goodchild, Michael F.; Maguire, David W.; Rhind (2005), Geographic Information Systems and Science (2nd ed.), Chichester: Wiley, ISBN 047087001X
- [24] Slocum, T. (2003). Thematic Cartography and Geographic Visualization. Upper Saddle River, New Jersey: Prentice Hall. ISBN 0-130-35123-7.
- [25] Kraak, Menno-Jan and Ormeling, Ferjan (2002). Cartography: Visualization of Spatial Data. Prentice Hall. ISBN 0-130-88890-7.
- [26] MacEachren, A.M. (1994). Some Truth with Maps: A Primer on Symbolization & Design. University Park: The Pennsylvania State University.
- [27] Bender, B. (1999). Subverting the Western Gaze: mapping alternative worlds. The Archaeology and Anthropology of Landscape: Shaping your landscape (eds) P.J. Ucko & R. Layton. London: Routledge.
- [28] Monmonier, Mark (1993). Mapping It Out. Chicago: University of Chicago Press. ISBN.
- [29] "GIS software training from Esri in India". NIIT Technology News. <http://www.prdomain.com/companies/N/NIIT/newsreleases/20055521656.htm>. Retrieved 2008-04-03.
- [30] "COTS GIS: The Value of a Commercial Geographic Information System". www.esri.com. <http://www.esri.com/library/whitepapers/pdfs/cots-gis.pdf>. Retrieved 2010-01-30.
- [31] "GIS Software Applications". gislounge.com. <http://gislounge.com/gis-software-applications/>. Retrieved 2010-01-30.
- [32] <http://apb.directionsmag.com/archives/7575-Esri-Transitioning-Pronunciation-of-its-Name.html>
- [33] "GIS in Our World". Esri Website: Company Facts. <http://www.esri.com/company/about/facts.html>. Retrieved 2008-04-03.
- [34] Matteo Luccio (2007-06-22). "Esri International User Conference". GIS Monitor Newsletter. GIS Monitor. <http://www.gismonitor.com/news/newsletter/archive/archives.php?issue=20070622>. Retrieved 2008-01-09.
- [35] Guy McCarthy; George Watson (2006-06-30). "Esri verifies company targeted by subpoena". The San Bernardino Sun. http://www.sbsun.com/onlineextras/ci_3996395. Retrieved May 25, 2009.
- [36] http://en.wikipedia.org/wiki/Geographic_information_system
- [37] <http://www.webopedia.com/TERM/G/GIS.html>
- [38] <http://www.gisdevelopment.net/tutorials/tuman006pf.htm>
- [39] <http://searchsqlserver.techtarget.com/definition/GIS>
- [40] http://www.agt.bme.hu/public_e/funcint/funcint.html
- [41] <http://www.ika.ethz.ch/schneider/Publications/SchneiderOttawa99.pdf>
- [42] <http://gisandscience.com/2009/09/14/top-five-benefits-of-gis/>
- [43] <http://maic.jmu.edu/sic/gis/components.htm>
- [44] http://www.sfu.ca/rdl/GIS/tour/comp_gis.html
- [45] <http://www.mapsofworld.com/gis/components.html>
- [46] <http://gis.nic.in/gisprimer/functions.html>
- [47] <http://maps.unomaha.edu/Peterson/gis/notes/GISAnal2.html>
- [48] http://healthcybermap.org/HGeo/pg2_3.htm
- [49] <http://mapzone.ordnancesurvey.co.uk/mapzone/giszone/english/gisfocus/page3.htm>
- [50] http://ittraining.iu.edu/workshops/schedule_by_group.aspx?group=15
- [51] http://www.sul.stanford.edu/depts/gis/project_design.html
- [51] <http://www.gis.com/content/data-types-and-models>
- [51] <http://www.gis.com/content/gis-data>
- [52] http://bgis.sanbi.org/gis-primer/page_14.htm
- [53] http://en.mimi.hu/gis/data_model.html
- [54] [http://en.wikipedia.org/wiki/Data_model_\(GIS\)](http://en.wikipedia.org/wiki/Data_model_(GIS))
- [55] http://bgis.sanbi.org/gis-primer/page_13.htm
- [56] <http://www.gis.com/content/data-types-and-models>
- [57] http://bgis.sanbi.org/GIS-primer/page_15.htm
- [58] <http://www.spatialanalysisonline.com/output/html/SpatialDataModelsandMethods.html>
- [59] <http://www.ncgia.ucsb.edu/giscc/units/u045/u045.html>
- [60] http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Topology_in_mathematics.html
- [61] <https://engineering.purdue.edu/~engelb/ncgia96/engel.html>
- [62] <http://www.dcsme.ae/>
- [63] <http://www.dcsa.com.au/09apr/index.php>
- [64] <http://www.datacapture.co.uk/>
- [65] http://en.wikipedia.org/wiki/Operational_data_store
- [66] <http://www.vdstech.com/aspmap.htm>
- [67] <http://www.freedownloadcenter.com/Programming/ActiveX/AspMap.html>
- [68] <http://www.opengeospatial.org/ogc/members>
- [69] http://www.cdrgroup.co.uk/index.htm?sales_mi_prod_planaccess.htm
- [70] <http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3>
- [71] https://njgin.state.nj.us/NJ_NJGINExplorer/IW.jsp
- [72] <http://nationalmap.gov/gio/standards>
- [73] http://gis.ednet.ns.ca/gis_uses_in_US.htm

TABLE I. COMPARISON BETWEEN REAL WORLD CRITERIONS AND ITS GIS REPRESENTATION

Real world geographical criterions	GIS representation layer
Roads Network	Line
Road Intersection	Node
Rivers	Line
Garden	Polygon
Public buildings	Polygon
Train Stations/ Ports	Polygon
Water Net Stations	Line
Gas Network Stations	Line
Sewers	Line

TABLE II. DIFFERENCE BETWEEN TRADITIONAL MANUAL AND NOVEL GIS E-SERVICE

Factors	Traditional manual service	GIS E-service
Time	Long time	Less time
Effort	Very High	Very Low

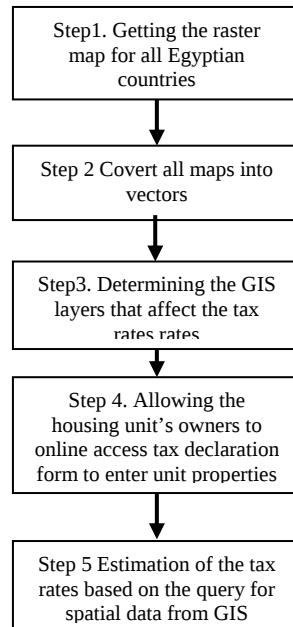


Figure 1. Block diagram for the steps of the novel e-services

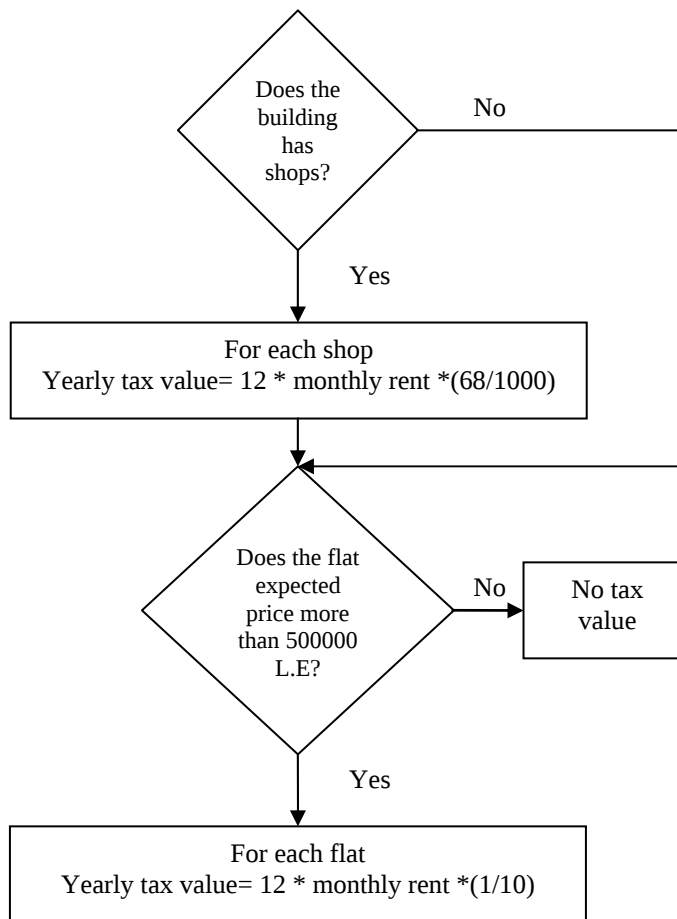


Figure 2. Flow chart of the computational method for housing tax rates

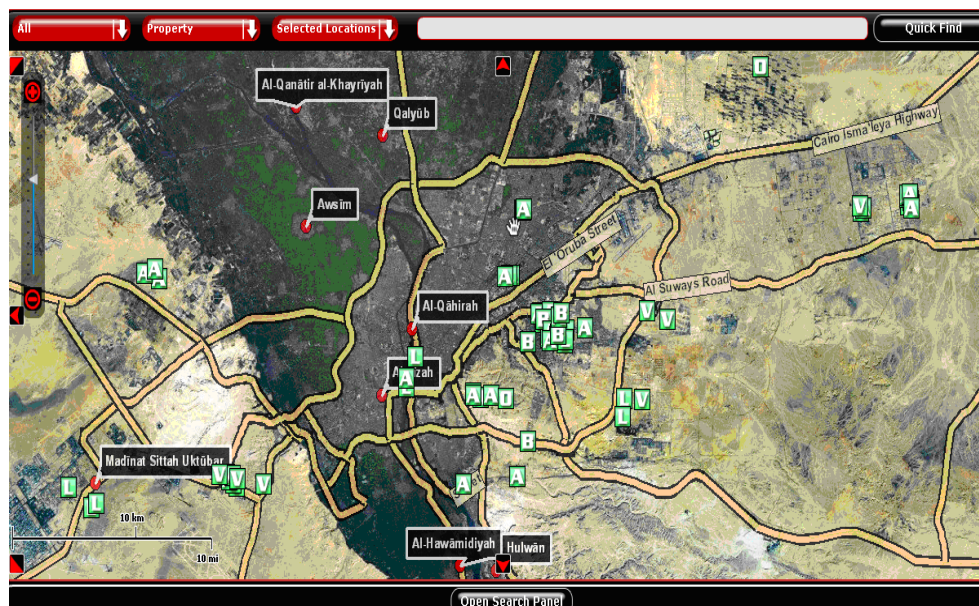


Figure 3. Interface of the proposed GIS e-service

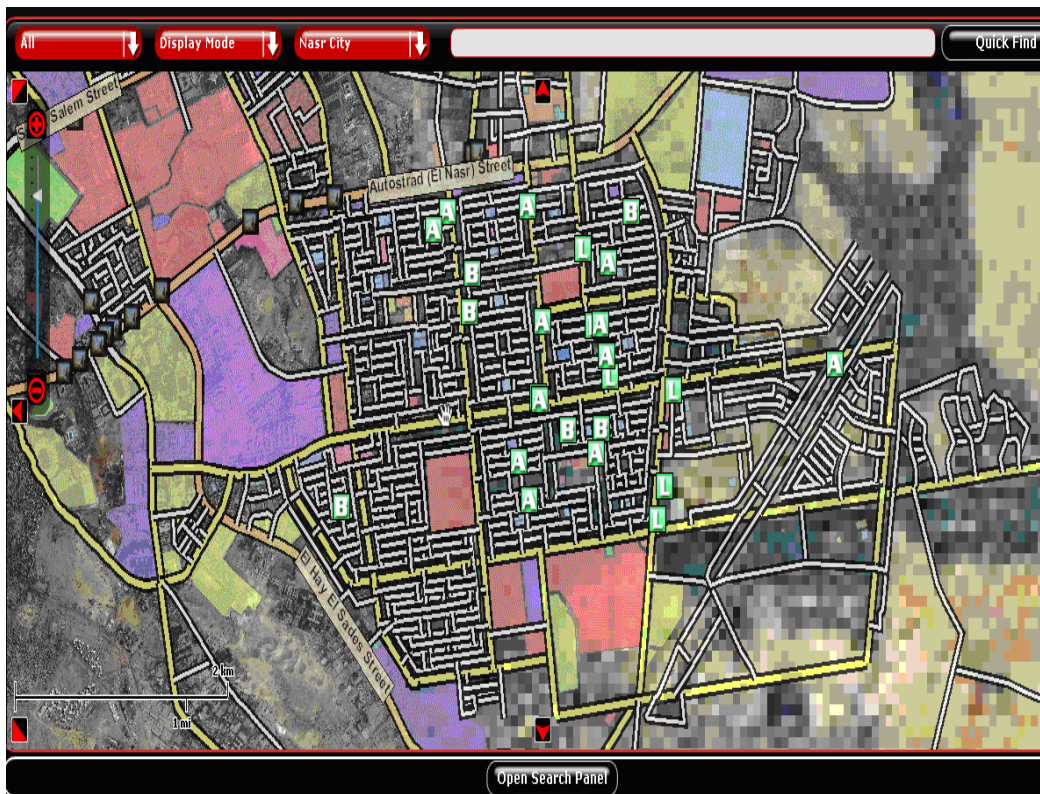


Figure 4. Interface for final GIS layers

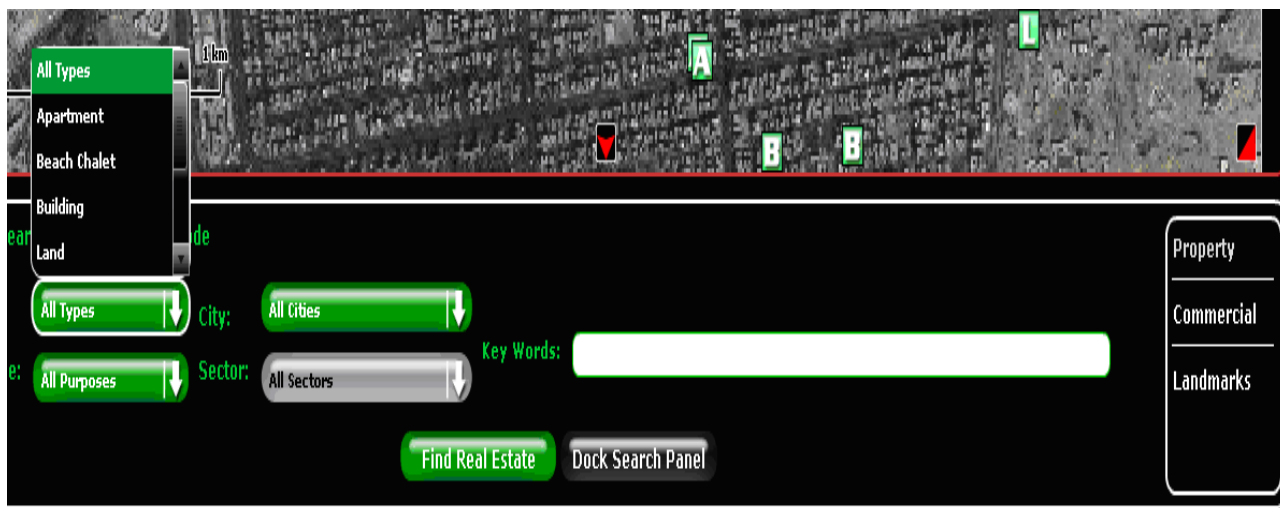


Figure 5. Query Interface for the committee

Property Details

Listing ID: 38

About the Owner / Submitter

Name: El Mostafa Email: algazwy2008@yahoo.com

Mobile Phone: 0126434431 Office Phone: Not Supplied

About the Property/Real Estate

Sector: Nasr City City: Cairo

Property Type: Apartment Property Purpose: For Sale

Price: EGP 900,000 Property Address: Next to the Serag City Mall

Description:
240 square meters, 7th floor

OK

Figure 6. Interface of main property details that entered by the owner of the unit

Measurement Results

Latitude :31.347
Longitude :30.055

OK

Figure 7. Results of committee's query about the latitude and longitude of the investigated building

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore, Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr. Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGILIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr. C. Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.) / Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V. Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr. Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia

Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Mr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Mr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan
Prof. Syed S. Rizvi, University of Bridgeport, USA
Mr. Shahbaz Pervez Chattha, University of Engineering and Technology Taxila, Pakistan
Dr. Shishir Kumar, Jaypee University of Information Technology, Wakanaghat (HP), India
Mr. Shahid Mumtaz, Portugal Telecommunication, Instituto de Telecomunicações (IT) , Aveiro, Portugal
Mr. Rajesh K Shukla, Corporate Institute of Science & Technology Bhopal M P
Dr. Poonam Garg, Institute of Management Technology, India
Mr. S. Mehta, Inha University, Korea
Mr. Dilip Kumar S.M, University Visvesvaraya College of Engineering (UVCE), Bangalore University,
Bangalore
Prof. Malik Sikander Hayat Khiyal, Fatima Jinnah Women University, Rawalpindi, Pakistan
Dr. Virendra Gomase , Department of Bioinformatics, Padmashree Dr. D.Y. Patil University
Dr. Irraivan Elamvazuthi, University Technology PETRONAS, Malaysia
Mr. Saqib Saeed, University of Siegen, Germany
Mr. Pavan Kumar Gorakavi, IPMA-USA [YC]
Dr. Ahmed Nabih Zaki Rashed, Menoufia University, Egypt
Prof. Shishir K. Shandilya, Rukmani Devi Institute of Science & Technology, India
Mrs.J.Komala Lakshmi, SNR Sons College, Computer Science, India
Mr. Muhammad Sohail, KUST, Pakistan
Dr. Manjaiah D.H, Mangalore University, India
Dr. S Santhosh Baboo, D.G.Vaishnav College, Chennai, India
Prof. Dr. Mokhtar Beldjehem, Sainte-Anne University, Halifax, NS, Canada
Dr. Deepak Laxmi Narasimha, Faculty of Computer Science and Information Technology, University of
Malaya, Malaysia
Prof. Dr. Arunkumar Thangavelu, Vellore Institute Of Technology, India
Mr. M. Azath, Anna University, India
Mr. Md. Rabiul Islam, Rajshahi University of Engineering & Technology (RUET), Bangladesh
Mr. Aos Alaa Zaidan Ansaef, Multimedia University, Malaysia
Dr Suresh Jain, Professor (on leave), Institute of Engineering & Technology, Devi Ahilya University, Indore
(MP) India,
Dr. Mohammed M. Kadhum, Universiti Utara Malaysia
Mr. Hanumanthappa. J. University of Mysore, India
Mr. Syed Ishtiaque Ahmed, Bangladesh University of Engineering and Technology (BUET)
Mr Akinola Solomon Olalekan, University of Ibadan, Ibadan, Nigeria

Mr. Santosh K. Pandey, Department of Information Technology, The Institute of Chartered Accountants of India

Dr. P. Vasant, Power Control Optimization, Malaysia

Dr. Petr Ivankov, Automatika - S, Russian Federation

Dr. Utkarsh Seetha, Data Infosys Limited, India

Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal

Dr. (Mrs) Padmavathi Ganapathi, Avinashilingam University for Women, Coimbatore

Assist. Prof. A. Neela madheswari, Anna university, India

Prof. Ganesan Ramachandra Rao, PSG College of Arts and Science, India

Mr. Kamanashis Biswas, Daffodil International University, Bangladesh

Dr. Atul Gonsai, Saurashtra University, Gujarat, India

Mr. Angkoon Phinyomark, Prince of Songkla University, Thailand

Mrs. G. Nalini Priya, Anna University, Chennai

Dr. P. Subashini, Avinashilingam University for Women, India

Assoc. Prof. Vijay Kumar Chakka, Dhirubhai Ambani IICT, Gandhinagar ,Gujarat

Mr Jitendra Agrawal, : Rajiv Gandhi Proudhyogiki Vishwavidyalaya, Bhopal

Mr. Vishal Goyal, Department of Computer Science, Punjabi University, India

Dr. R. Baskaran, Department of Computer Science and Engineering, Anna University, Chennai

Assist. Prof, Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India

Dr. Jamal Ahmad Dargham, School of Engineering and Information Technology, Universiti Malaysia Sabah

Mr. Nitin Bhatia, DAV College, India

Dr. Dhavachelvan Ponnurangam, Pondicherry Central University, India

Dr. Mohd Faizal Abdollah, University of Technical Malaysia, Malaysia

Assist. Prof. Sonal Chawla, Panjab University, India

Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India

Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia

Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia

Professor Dr. Sabu M. Thampi, .B.S Institute of Technology for Women, Kerala University, India

Mr. Noor Muhammed Nayeem, Université Lumière Lyon 2, 69007 Lyon, France

Dr. Himanshu Aggarwal, Department of Computer Engineering, Punjabi University, India

Prof R. Naidoo, Dept of Mathematics/Center for Advanced Computer Modelling, Durban University of Technology, Durban,South Africa

Prof. Mydhili K Nair, M S Ramaiah Institute of Technology(M.S.R.I.T), Affiliated to Visweswaraiah Technological University, Bangalore, India

M. Prabu, Adhiyamaan College of Engineering/Anna University, India

Mr. Swakkhar Shatabda, Department of Computer Science and Engineering, United International University, Bangladesh

Dr. Abdur Rashid Khan, ICIT, Gomal University, Dera Ismail Khan, Pakistan

Mr. H. Abdul Shabeer, I-Nautix Technologies,Chennai, India

Dr. M. Aramudhan, Perunthalaivar Kamarajar Institute of Engineering and Technology, India

Dr. M. P. Thapliyal, Department of Computer Science, HNB Garhwal University (Central University), India
Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
Mr. Zeashan Hameed Khan, : Université de Grenoble, France
Prof. Anil K Ahlawat, Ajay Kumar Garg Engineering College, Ghaziabad, UP Technical University, Lucknow
Mr. Longe Olumide Babatope, University Of Ibadan, Nigeria
Associate Prof. Raman Maini, University College of Engineering, Punjabi University, India
Dr. Maslin Masrom, University Technology Malaysia, Malaysia
Sudipta Chattopadhyay, Jadavpur University, Kolkata, India
Dr. Dang Tuan NGUYEN, University of Information Technology, Vietnam National University - Ho Chi Minh City
Dr. Mary Lourde R., BITS-PILANI Dubai , UAE
Dr. Abdul Aziz, University of Central Punjab, Pakistan
Mr. Karan Singh, Gautam Budtha University, India
Mr. Avinash Pokhriyal, Uttar Pradesh Technical University, Lucknow, India
Associate Prof Dr Zuraini Ismail, University Technology Malaysia, Malaysia
Assistant Prof. Yasser M. Alginahi, College of Computer Science and Engineering, Taibah University, Madinah Munawwarah, KSA
Mr. Dakshina Ranjan Kisku, West Bengal University of Technology, India
Mr. Raman Kumar, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India
Associate Prof. Samir B. Patel, Institute of Technology, Nirma University, India
Dr. M.Munir Ahamed Rabbani, B. S. Abdur Rahman University, India
Asst. Prof. Koushik Majumder, West Bengal University of Technology, India
Dr. Alex Pappachen James, Queensland Micro-nanotechnology center, Griffith University, Australia
Assistant Prof. S. Hariharan, B.S. Abdur Rahman University, India
Asst Prof. Jasmine. K. S, R.V.College of Engineering, India
Mr Naushad Ali Mamode Khan, Ministry of Education and Human Resources, Mauritius
Prof. Mahesh Goyani, G H Patel Collge of Engg. & Tech, V.V.N, Anand, Gujarat, India
Dr. Mana Mohammed, University of Tlemcen, Algeria
Prof. Jatinder Singh, Universal Institutiion of Engg. & Tech. CHD, India
Mrs. M. Anandhavalli Gauthaman, Sikkim Manipal Institute of Technology, Majitar, East Sikkim
Dr. Bin Guo, Institute Telecom SudParis, France
Mrs. Maleika Mehr Nigar Mohamed Heenaye-Mamode Khan, University of Mauritius
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Mr. V. Bala Dhandayuthapani, Mekelle University, Ethiopia
Dr. Irfan Syamsuddin, State Polytechnic of Ujung Pandang, Indonesia
Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
Mr. Ravi Chandiran, Zagro Singapore Pte Ltd. Singapore
Mr. Milindkumar V. Sarode, Jawaharlal Darda Institute of Engineering and Technology, India
Dr. Shamimul Qamar, KSJ Institute of Engineering & Technology, India
Dr. C. Arun, Anna University, India

Assist. Prof. M.N.Birje, Basaveshwar Engineering College, India
Prof. Hamid Reza Naji, Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran
Assist. Prof. Debasis Giri, Department of Computer Science and Engineering, Haldia Institute of Technology
Subhabrata Barman, Haldia Institute of Technology, West Bengal
Mr. M. I. Lali, COMSATS Institute of Information Technology, Islamabad, Pakistan
Dr. Feroz Khan, Central Institute of Medicinal and Aromatic Plants, Lucknow, India
Mr. R. Nagendran, Institute of Technology, Coimbatore, Tamilnadu, India
Mr. Amnach Khawne, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, Thailand
Dr. P. Chakrabarti, Sir Padampat Singhanian University, Udaipur, India
Mr. Nafiz Imtiaz Bin Hamid, Islamic University of Technology (IUT), Bangladesh.
Shahab-A. Shamsirband, Islamic Azad University, Chalous, Iran
Prof. B. Priestly Shan, Anna Univeristy, Tamilnadu, India
Venkatramreddy Velma, Dept. of Bioinformatics, University of Mississippi Medical Center, Jackson MS USA
Akshi Kumar, Dept. of Computer Engineering, Delhi Technological University, India
Dr. Umesh Kumar Singh, Vikram University, Ujjain, India
Mr. Serguei A. Mokhov, Concordia University, Canada
Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia
Dr. Awadhesh Kumar Sharma, Madan Mohan Malviya Engineering College, India
Mr. Syed R. Rizvi, Analytical Services & Materials, Inc., USA
Dr. S. Karthik, SNS College of Technology, India
Mr. Syed Qasim Bukhari, CIMET (Universidad de Granada), Spain
Mr. A.D.Potgantwar, Pune University, India
Dr. Himanshu Aggarwal, Punjabi University, India
Mr. Rajesh Ramachandran, Naipunya Institute of Management and Information Technology, India
Dr. K.L. Shunmuganathan, R.M.K Engg College, Kavaraipeitai, Chennai
Dr. Prasant Kumar Pattnaik, KIST, India.
Dr. Ch. Aswani Kumar, VIT University, India
Mr. Ijaz Ali Shoukat, King Saud University, Riyadh KSA
Mr. Arun Kumar, Sir Padam Pat Singhanian University, Udaipur, Rajasthan
Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
Mr. Mohd Zaki Bin Mas'ud, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
Prof. Dr. R. Geetharamani, Dept. of Computer Science and Eng., Rajalakshmi Engineering College, India
Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
Dr. S. Abdul Khader Jilani, University of Tabuk, Tabuk, Saudi Arabia
Mr. Syed Jamal Haider Zaidi, Bahria University, Pakistan
Dr. N. Devarajan, Government College of Technology, Coimbatore, Tamilnadu, INDIA
Mr. R. Jagadeesh Kannan, RMK Engineering College, India
Mr. Deo Prakash, Shri Mata Vaishno Devi University, India
Mr. Mohammad Abu Naser, Dept. of EEE, IUT, Gazipur, Bangladesh

Assist. Prof. Prasun Ghosal, Bengal Engineering and Science University, India
Mr. Md. Golam Kaosar, School of Engineering and Science, Victoria University, Melbourne City, Australia
Mr. R. Mahammad Shafi, Madanapalle Institute of Technology & Science, India
Dr. F.Sagayaraj Francis, Pondicherry Engineering College, India
Dr. Ajay Goel, HIET , Kaithal, India
Mr. Nayak Sunil Kashibarao, Bahirji Smarak Mahavidyalaya, India
Mr. Suhas J Manangi, Microsoft India
Dr. Kalyankar N. V., Yeshwant Mahavidyalaya, Nanded , India
Dr. K.D. Verma, S.V. College of Post graduate studies & Research, India
Dr. Amjad Rehman, University Technology Malaysia, Malaysia
Mr. Rachit Garg, L K College, Jalandhar, Punjab
Mr. J. William, M.A.M college of Engineering, Trichy, Tamilnadu, India
Prof. Jue-Sam Chou, Nanhua University, College of Science and Technology, Taiwan
Dr. Thorat S.B., Institute of Technology and Management, India
Mr. Ajay Prasad, Sir Padampat Singhanian University, Udaipur, India
Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology & Science, India
Mr. Syed Rafiul Hussain, Ahsanullah University of Science and Technology, Bangladesh
Mrs Fazeela Tunnisa, Najran University, Kingdom of Saudi Arabia
Mrs Kavita Taneja, Maharishi Markandeshwar University, Haryana, India
Mr. Maniyar Shiraz Ahmed, Najran University, Najran, KSA
Mr. Anand Kumar, AMC Engineering College, Bangalore
Dr. Rakesh Chandra Gangwar, Beant College of Engg. & Tech., Gurdaspur (Punjab) India
Dr. V V Rama Prasad, Sree Vidyanikethan Engineering College, India
Assist. Prof. Neetesh Kumar Gupta, Technocrats Institute of Technology, Bhopal (M.P.), India
Mr. Ashish Seth, Uttar Pradesh Technical University, Lucknow ,UP India
Dr. V V S S S Balaram, Sreenidhi Institute of Science and Technology, India
Mr Rahul Bhatia, Lingaya's Institute of Management and Technology, India
Prof. Niranjana Reddy. P, KITS , Warangal, India
Prof. Rakesh. Lingappa, Vijetha Institute of Technology, Bangalore, India
Dr. Mohammed Ali Hussain, Nimra College of Engineering & Technology, Vijayawada, A.P., India
Dr. A.Srinivasan, MNM Jain Engineering College, Rajiv Gandhi Salai, Thorapakkam, Chennai
Mr. Rakesh Kumar, M.M. University, Mullana, Ambala, India
Dr. Lena Khaled, Zarqa Private University, Aman, Jordan
Ms. Supriya Kapoor, Patni/Lingaya's Institute of Management and Tech., India
Dr. Tossapon Boongoen , Aberystwyth University, UK
Dr . Bilal Alatas, Firat University, Turkey
Assist. Prof. Jyoti Praaksh Singh , Academy of Technology, India
Dr. Ritu Soni, GNG College, India
Dr . Mahendra Kumar , Sagar Institute of Research & Technology, Bhopal, India.
Dr. Binod Kumar, India

Dr. Muzhir Shaban Al-Ani, Amman Arab University Amman – Jordan
Dr. T.C. Manjunath , ATRIA Institute of Tech, India
Mr. Muhammad Zakarya, COMSATS Institute of Information Technology (CIIT), Pakistan
Assist. Prof. Harmunish Taneja, M. M. University, India
Dr. Chitra Dhawale , SICSIR, Model Colony, Pune, India
Mrs Sankari Muthukaruppan, Nehru Institute of Engineering and Technology, Anna University, India
Mr. Aaqif Afzaal Abbasi, National University Of Sciences And Technology, Islamabad
Prof. Ashutosh Kumar Dubey, Trinity Institute of Technology and Research Bhopal, India
Mr. G. Appasami, Dr. Pauls Engineering College, India
Mr. M Yasin, National University of Science and Tech, karachi (NUST), Pakistan
Mr. Yaser Miaji, University Utara Malaysia, Malaysia
Mr. Shah Ahsanul Haque, International Islamic University Chittagong (IIUC), Bangladesh
Prof. (Dr) Syed Abdul Sattar, Royal Institute of Technology & Science, India
Dr. S. Sasikumar, Roever Engineering College
Assist. Prof. Monit Kapoor, Maharishi Markandeshwar University, India
Mr. Nwaocha Vivian O, National Open University of Nigeria
Dr. M. S. Vijaya, GR Govindarajulu School of Applied Computer Technology, India
Assist. Prof. Chakresh Kumar, Manav Rachna International University, India
Mr. Kunal Chadha , R&D Software Engineer, Gemalto, Singapore
Mr. Pawan Jindal, Jaypee University of Engineering and Technology, India
Mr. Mueen Uddin, Universiti Teknologi Malaysia, UTM , Malaysia
Dr. Dhuha Basheer abdullah, Mosul university, Iraq
Mr. S. Audithan, Annamalai University, India
Prof. Vijay K Chaudhari, Technocrats Institute of Technology , India
Associate Prof. Mohd Ilyas Khan, Technocrats Institute of Technology , India
Dr. Vu Thanh Nguyen, University of Information Technology, HoChiMinh City, VietNam
Assist. Prof. Anand Sharma, MITS, Lakshmangarh, Sikar, Rajasthan, India
Prof. T V Narayana Rao, HITAM Engineering college, Hyderabad
Mr. Deepak Gour, Sir Padampat Singhania University, India
Assist. Prof. Amutharaj Joyson, Kalasalingam University, India
Mr. Ali Balador, Islamic Azad University, Iran
Mr. Mohit Jain, Maharaja Surajmal Institute of Technology, India
Mr. Dilip Kumar Sharma, GLA Institute of Technology & Management, India
Dr. Debojyoti Mitra, Sir padampat Singhania University, India
Dr. Ali Dehghantanha, Asia-Pacific University College of Technology and Innovation, Malaysia
Mr. Zhao Zhang, City University of Hong Kong, China
Prof. S.P. Setty, A.U. College of Engineering, India
Prof. Patel Rakeshkumar Kantilal, Sankalchand Patel College of Engineering, India
Mr. Biswajit Bhowmik, Bengal College of Engineering & Technology, India
Mr. Manoj Gupta, Apex Institute of Engineering & Technology, India

Assist. Prof. Ajay Sharma, Raj Kumar Goel Institute Of Technology, India
Assist. Prof. Ramveer Singh, Raj Kumar Goel Institute of Technology, India
Dr. Hanan Elazhary, Electronics Research Institute, Egypt
Dr. Hosam I. Faiq, USM, Malaysia
Prof. Dipti D. Patil, MAEER's MIT College of Engg. & Tech, Pune, India
Assist. Prof. Devendra Chack, BCT Kumaon engineering College Dwarahat Almora, India
Prof. Manpreet Singh, M. M. Engg. College, M. M. University, India
Assist. Prof. M. Sadiq ali Khan, University of Karachi, Pakistan
Mr. Prasad S. Halgaonkar, MIT - College of Engineering, Pune, India
Dr. Imran Ghani, Universiti Teknologi Malaysia, Malaysia
Prof. Varun Kumar Kakar, Kumaon Engineering College, Dwarahat, India
Assist. Prof. Nisheeth Joshi, Apaji Institute, Banasthali University, Rajasthan, India
Associate Prof. Kunwar S. Vaisla, VCT Kumaon Engineering College, India
Prof Anupam Choudhary, Bhilai School Of Engg., Bhilai (C.G.), India
Mr. Divya Prakash Shrivastava, Al Jabal Al garbi University, Zawya, Libya
Associate Prof. Dr. V. Radha, Avinashilingam Deemed university for women, Coimbatore.
Dr. Kasarapu Ramani, JNT University, Anantapur, India
Dr. Anuraag Awasthi, Jayoti Vidyapeeth Womens University, India
Dr. C G Ravichandran, R V S College of Engineering and Technology, India
Dr. Mohamed A. Deriche, King Fahd University of Petroleum and Minerals, Saudi Arabia
Mr. Abbas Karimi, Universiti Putra Malaysia, Malaysia
Mr. Amit Kumar, Jaypee University of Engg. and Tech., India
Dr. Nikolai Stoianov, Defense Institute, Bulgaria
Assist. Prof. S. Ranichandra, KSR College of Arts and Science, Tiruchencode
Mr. T.K.P. Rajagopal, Diamond Horse International Pvt Ltd, India
Dr. Md. Ekramul Hamid, Rajshahi University, Bangladesh
Mr. Hemanta Kumar Kalita , TATA Consultancy Services (TCS), India
Dr. Messaouda Azzouzi, Ziane Achour University of Djelfa, Algeria
Prof. (Dr.) Juan Jose Martinez Castillo, "Gran Mariscal de Ayacucho" University and Acantelys research Group, Venezuela
Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India
Dr. Babak Bashari Rad, University Technology of Malaysia, Malaysia
Mr. B. Muthu Kumar, Kathir College Of Engineering, Coimbatore
Dr. Nighat Mir, Effat University, Saudi Arabia
Prof. (Dr.) G.M.Nasira, Sasurie College of Engineering, India
Mr. Varun Mittal, Gemalto Pte Ltd, Singapore
Assist. Prof. Mrs P. Banumathi, Kathir College Of Engineering, Coimbatore
Assist. Prof. Quan Yuan, University of Wisconsin-Stevens Point, US
Dr. Pranam Paul, Narula Institute of Technology, Agarpara, West Bengal, India
Assist. Prof. J. Ramkumar, V.L.B Janakiammal college of Arts & Science, India

Mr. P. Sivakumar, Anna university, Chennai, India
Mr. Md. Humayun Kabir Biswas, King Khalid University, Kingdom of Saudi Arabia
Mr. Mayank Singh, J.P. Institute of Engg & Technology, Meerut, India
HJ. Kamaruzaman Jusoff, Universiti Putra Malaysia
Mr. Nikhil Patrick Lobo, CADES, India
Mr. Amit Wason, Rayat-Bahra Institute of Engineering & Boi-Technology, India
Dr. Rajesh Shrivastava, Govt. Benazir Science & Commerce College, Bhopal, India
Assist. Prof. Vishal Bharti, DCE, Gurgaon
Mrs. Sunita Bansal, Birla Institute of Technology & Science, India
Dr. R. Sudhakar, Dr.Mahalingam college of Engineering and Technology, India
Dr. Amit Kumar Garg, Shri Mata Vaishno Devi University, Katra(J&K), India
Assist. Prof. Raj Gaurang Tiwari, AZAD Institute of Engineering and Technology, India
Mr. Hamed Taherdoost, Tehran, Iran
Mr. Amin Daneshmand Malayeri, YRC, IAU, Malayer Branch, Iran
Mr. Shantanu Pal, University of Calcutta, India
Dr. Terry H. Walcott, E-Promag Consultancy Group, United Kingdom
Dr. Ezekiel U OKIKE, University of Ibadan, Nigeria
Mr. P. Mahalingam, Caledonian College of Engineering, Oman
Dr. Mahmoud M. A. Abd Ellatif, Mansoura University, Egypt
Prof. Kunwar S. Vaisla, BCT Kumaon Engineering College, India
Prof. Mahesh H. Panchal, Kalol Institute of Technology & Research Centre, India
Mr. Muhammad Asad, University of Engineering and Technology Taxila, Pakistan
Mr. AliReza Shams Shafigh, Azad Islamic university, Iran
Prof. S. V. Nagaraj, RMK Engineering College, India
Mr. Ashikali M Hasan, Senior Researcher, CelNet security, India
Dr. Adnan Shahid Khan, University Technology Malaysia, Malaysia
Mr. Prakash Gajanan Burade, Nagpur University/ITM college of engg, Nagpur, India
Dr. Jagdish B.Helonde, Nagpur University/ITM college of engg, Nagpur, India
Professor, Doctor BOUHORMA Mohammed, Univertsity Abdelmalek Essaadi, Morocco
Mr. K. Thirumalaivasan, Pondicherry Engg. College, India

CALL FOR PAPERS
International Journal of Computer Science and Information Security
IJCSIS 2011
ISSN: 1947-5500
<http://sites.google.com/site/ijcsis/>

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc.), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on

its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embedded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at <http://sites.google.com/site/ijcsis/authors-notes> .



© IJCSIS PUBLICATION 2011
ISSN 1947 5500